Metagenomics Workshop

Lead by Regina Lamendella, Juniata College
lamendella@juniata.edu
814-641-3553

Acknowledgements: I would like to thank Ms. Erin McClure for her help in developing many of these tutorials and for her help in preparing this document

Table of Contents

16S rRNA Gene and ITS Library Preparation

A. Background
B. Library Preparation
   1. 16S rRNA gene Illumina tag (itag) PCR
   2. ITS Illumina tag (itag) PCR
C. Check PCR Amplification
   1. Pool replicate samples
   2. E-gel electrophoresis
   3. DNA quantification with the Qubit fluorometer
      a. Introduction
      b. Materials
      c. Protocol
D. Quality Check Libraries
   1. Pool samples
   2. Gel electrophoresis
   3. QIAquick gel purification
   4. Bioanalyzer
      a. Introduction
      b. Agilent high sensitivity DNA assay protocol
      c. Interpreting Bioanalyzer results
E. References and suggested reading

Bioinformatics

A. Background
B. Unix/Linux tutorial
C. Introduction to QIIME
D. Installing the QIIME VirtualBox image
E. QIIME 16S workflow
   1. Conventions
   2. Flowchart
   3. Metadata
   4. Extract compressed files
   5. Split libraries workaround
   6. OTU table picking

7. Initial analyses
    a. OTU table statistics
    b. Clean OTU table
    c. Summarize taxa
8. Diversity analyses
    a. Alpha diversity
    b. Beta diversity
9. Statistical analyses
    a. OTU category significance
    b. Compare categories
    c. Compare alpha diversity
10. Heatmaps

F. QIIME ITS workflow
1. Obtain tutorial files
2. OTU picking

G. References and suggested reading

APPENDIX

A. Helpful links
B. Additional protocols/scripts
1. Purification by SPRI beads
2. Splitting libraries – the traditional method

C. Other software
1. Other ways to analyze metagenomes
2. Installing R
3. Proprietary software for data analysis

D. Computing
1. Troubleshooting error messages
2. Connecting to Juniata's HHMI cluster
3. IPython Notebook
4. Bash scripting

E. References

MODULE 1: PREPARATION OF MICROBIAL SAMPLES FOR HIGH-THROUGHPUT SEQUENCING

After this module you will be able to show your children how to do this…I promise!

Background

The term 'metagneomics' was originally coined by Jo Handelsman in the late 1990s and is currently defined as the application of modern genomics techniques to the study of microbial communities directly in their natural environments". The culture-independent molecular techniques have allowed microbiologists to tap into the vast microbial diversity of our world. Recently, massively parallel, high-throughput sequencing (HTS) has enabled taxonomic profiling of microbial communities to become cost-effective and informative. Initiatives such as the Earth Microbiome Project, the Hospital Microbiome Project, the Human Microbiome Project, and others are consortia tasked with uncovering the distribution of microorganisms within us and our world. Many efforts have focused on probing regions of the ribosomal RNA operon as a method for 'telling us who is there in our sample". The rRNA operon, contains genes encoding structural and functional portions of the ribosome. This operon contains both highly conserved and highly variable regions, which allow specific regions to be simultaneously targeted and to distinguish taxa from one another. Microbiologists have relied upon DNA sequence information for microbial identification, based primarily on the gene encoding the small subunit RNA molecule of the ribosome (16S rRNA gene). Databases of rRNA sequence data can be used to design phylogenetically conserved probes that target both individual and closely related groups of microorganisms without cultivation. Some of the most well curated databases of 16S rRNA sequences include Greengenes, the Ribosomal Database Project, and ARB-Silva (see references section for links to these databases).

Figure 1. Structure of the rRNA operon in bacteria. Figure from Principles of Biochemistry 4th Edition Pearson Prentice Hall Inc. 2006

The ribosomal RNA genes (encoding 16S, 23S and 5S rRNAs) are typically linked together with tRNA molecules into operons that are coordinately transcribed to produce equimolar

quantities of each gene product (Figure 1). “Universal” primers can be used to amplify regions of the prokaryotic 16S rRNA gene. Approximately genus level taxonomic resolution can be achieved, depending on which variable region is amplified. Currently most widely used 16S rRNA region for high-throughput sequencing is the V4 region (Caporoso et al, 2010). A description of which regions are most useful for particular applications is described in Soergel et al.

Similarly, the internal transcribed spacer (ITS) regions of the rRNA gene in eukaryotes is used for taxonomic profiling in fungi. The ITS regions refer to pieces of non-functionaal RNA situated between structural ribosomal RNA on a common precursor transcript. Reading from the 5’ to 3’ direction, this precursor transcript contains the 5' external transcribed sequence (5' ETS), 18S rRNA, ITS1, 5.8S rRNA, ITS2, 28S rRNA and finally the 3'ETS. During rRNA maturation, ETS and ITS pieces are excised. The ITS varies greatly between fungal taxa, which has allowed it to be useful for determining which fungal taxa are present in a sample. This can be explained by the relatively low evolutionary pressure acting on such non-functional sequences. The ITS region is now perhaps the most widely sequenced DNA region in fungi (Peay et al., 2008). For the purposes of this module, we will amplify the ITS 1 region as described in McGuire et al.

Module Goals

- Participants will learn the structural and functional importance of the rRNA operon and its utility in studying microbial diversity.
- Participants will prepare bacterial and/or fungal sequencing libraries from DNA extracts by carrying out 16S rRNA and/or ITS library preparation using illumina tag PCR, E-gel electrophoresis, DNA quantification, gel purification, and library quality checking.
- Participants will also learn common issues associated with preparation of libraries and troubleshooting options.
- By the end of this module participants will have 16S rRNA gene and/or ITS libraries ready for submission for sequencing on the Illumina MiSeq platform.

Vision and Change core competencies addressed in this module

- Ability to apply the process of science by designing scientific process to understand microbial communities in their natural environments

- Ability to apply the process of science by developing problem-solving strategies to troubleshoot issues associated with PCR inhibition, and instrumentation.

- Ability to understand the relationship between science and society as participants will need to contextualize and convey how their project relates human health and/or the environment.

- Ability to tap into the interdisciplinary nature of science by applying physical and chemical principles of molecules to provide an in depth understanding of high-throughput sequencing technologies.

GCAT-SEEK sequencing requirements

Data for these libraries will be sequenced using the Illumina MiSeq platform. This technology currently yields up to 300 bp read lengths. Single end runs yield 12-15 million reads, while paired end read lengths yield 24-30 million reads. More information is available at:

- Video: http://www.youtube.com/watch?v=t0akxx8Dwsk
- Background Information: http://www.youtube.com/watch?v=t0akxx8Dwsk

Our prepared libraries for this workshop will be performed at the Dana Farber Sequencing Center. They offer full MiSeq runs for 1,000$ for educational research purposes. http://pcpgm.partners.org/research-services/sequencing/illumina

The BROAD Institute provides a great set of Illumina sequencing videos, which are really in depth and helpful. Visit: http://www.broadinstitute.org/scientific-community/science/platforms/genome-sequencing/broadillumina-genome-analyzer-boot-camp

Instrumentation and Supply requirements for this Module

1) Pipettes and tips
   For projects with more than 48 samples, multi-channel pipettes are helpful!

2) Qubit fluorometer- Life technologies, more information at: http://www.invitrogen.com/site/us/en/home/brands/Product-Brand/Qubit.html
   Note: The PicoGreen assay and a Spec reader is just as accurate as the Qubit 2.0 fluorometer. Nanodrop or specs that read 260/280 ratio can be used, but are not as accurate because other substances can absorb at the same wavelength as DNA and skew results.
3) Thermalcylcer- pretty much any one will do. At Juniata we use a mix of BIO-RAD's and MJ Research cyclers.
4) Electrophoresis unit- Any electrophoresis unit will work fine. We typically use between 1-2% gels for all applications. We stain our gels with GelStar GEL STAIN. Ethidium bromide is fine too. Any 1Kb ladder will suffice.
   For the initial check gel after PCR, we use the high-throughput E-gel system by Life Technologies to save time in the classroom. The gels are bufferless, precast, and

only take 12 minutes to run! More information on the Egel system can be found at http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/DNA-RNA-Purification-Analysis/Nucleic-Acid-Gel-Electrophoresis/E-Gel-Electrophoresis-System

5) PCR reagents: Qiagen HOTSTAR Kit is what we use for the PCR reagents in this module.
6) Primers used in this study were ordered from IDT. These primers were ordered in a 96-well format and were normalized to 3 nanomoles. The approximate cost is 28 cents per basepair. So each plate of primers costs roughly 2,000$ USD. Call your regional IDT rep and they will give you a sizeable discount. A list of the primers used in this study. More information can be found at:

Luckily we have tons of bacterial and fungal primers which we can send aliquots of directly to you, if needed.

A list of the primer constructs used in this module can be found in the Appendix.

7) Gel visualization apparatus. Any type of UV box with a camera adaptor will work.

8) Bioanalyzer 2100 and Expert software. More information on the bioanalyzer is available at: https://www.genomics.agilent.com/article.jsp?pageId=275

If you don't have a bioanalyzer, any sequencing facility can quality check your libraries for you for a small additional cost (roughly 100-150$/chip).

Table 1. List of reagents used in this module

Company	order number	Description	price 2013
Lonza	50535	GelStar GEL STAIN 10,000 0X (2 X 250uL)	$161.00
Qiagen	202602	HotStar HiFidelity Polymerase Kit (100)	$120.00
Qiagen	28604	MinElute Gel Extraction Kit (50)	$117.00
IDT	get quote	each primer is 68 bp x 28 cents/ base x 96 primers per plate	$1,827.84
Life Technologies	G7008-02	2% E-Gel® 96 Agarose	$216.00
Life Technologies	12373031	Egel low range ladder	$82.00
Agilent	5067-1504	Agilent DNA 1000 Kit (25 chips)	$739.00

Protocols

Some of protocols have been adapted from the Earth Microbiome Project. For further information please visit: http://www.earthmicrobiome.org/

A. Library Preparation

1. 16S rRNA gene Illumina tag (itag) PCR (*set up time 2 hours, runtime 3 hours*)

Illumina tag PCR amplification accomplishes two steps in one reaction. The desired region(s) of the 16S rRNA gene is amplified, which is typically required to obtain enough DNA for sequencing. By modifying the primer constructs to include the Illumina adapters and a unique barcode, the amplified region of the 16S rRNA gene can be identified in a pooled sample and the sample is prepared for sequencing with the Illumina MiSeq platform.

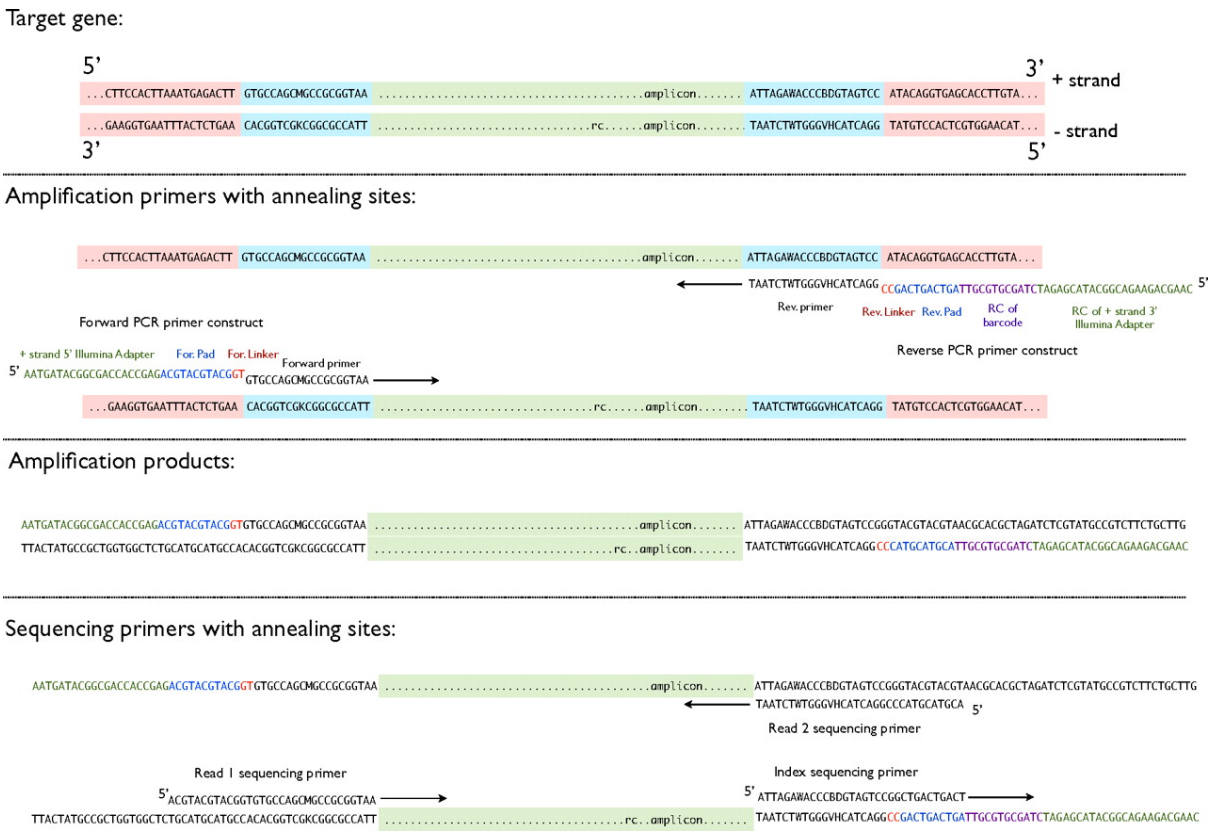


Figure 2. Protocol for barcoded Illumina sequencing. A target gene is identified, which in this case is the V4 region of the 16S rRNA gene. This region is PCR amplified using primer constructs with Illumina adapters, linker and pad sequences, and the forward/reverse primers themselves. The reverse primer construct contains an additional 12 bp barcode sequence. After PCR amplification, the target region is labeled with Illumina adapters and the barcode. The sequencing primers anneal and produce reads, while the index sequencing primer sequences the barcode. This information is used prior to analyzing the reads to demultiplex the sequences. See Caporaso et al (2011) for more information[1].

These PCR amplification protocols are based on the Earth Microbiome Project's list of standard protocols. (http://www.earthmicrobiome.org/emp-standard-protocols/16s/)

Reactions will be performed in duplicate. Record the PCR plate set up in the chart below or in the appropriate spreadsheet on the flash drive.

Reagent	[Initial]	Volume (µL)	[Final]
PCR grade H_2O		10.5	
HotStar Taq Plus MM kit	2X	12.5	1X
DNA template		1.0	
Forward Primer	10 µM	0.5	0.2 µM
Reverse primer	10 µM	0.5	0.2 µM
Total volume		**25.0**	

Thermocycling Conditions

1. 94°C for 3 min to denature the DNA
2. 94 °C for 45 s
3. 50 °C for 60 s
4. 72 °C for 90 s

} 35 cycles

5. 72 °C for 10 min for final extension
6. 4 °C HOLD

16S rRNA samples												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
A												
B												
C												
D												
E												
F												
G												
H												

2. ITS Illumina tag PCR

The ITS1-F and ITS2 primer pair[2] with appropriate Illumina adapters, pad and linker sequences, and 12 bp Golay barcodes is used to amplify the ITS-1 region. The PCR conditions are based on the Earth Microbiome Project standard amplification protocols and have been used in recent papers[3]. PCR reactions will be performed in duplicate.

ITS1-F CTTGGTCATTTAGAGGAAGTAA
ITS2 GCTGCGTTCTTCATCGATGC

Record the PCR plate set up on the chart below or in the appropriate spreadsheet on the flash drive.

Reagent	**[Initial]**	**Volume (µL)**	**[Final]**
DNA template		1.0	
HotStarTaq Plus MM kit	2X	12.5	1X
Forward Primer	10 µM	0.5	0.2 µM

Reverse primer	10 μM	0.5	0.2 μM
PCR grade H_2O		10.5	
Total volume		**25.0**	

Thermocycling Conditions

1. 94°C for 3 min to denature the DNA
2. 94 °C for 45 s
3. 50 °C for 60 s
4. 72 °C for 90 s

} 35 cycles

5. 72 °C for 10 min for final extension
6. 4 °C HOLD

ITS fungal samples												
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
A												
B												
C												
D												
E												
F												
G												
H												

C. Check PCR Amplification (1-2 hours)

1. Pooling the DNA (*30 mins- 1hour depending on the number of samples*)
Combine duplicate reactions into a single pool per sample. After combining, determine which PCR reactions were successful with the E-gel electrophoresis protocol.

2. E-Gel Electrophoresis (*15-30 mins for loading; 15 mins for runtime*)

E-Gels can be used to check the PCR product instead of traditional gel electrophoresis. We will only combine the successfully amplified samples for sequencing. We can also detect the presence of additional bands in the reactions, which may signal amplification of chloroplast DNA or excessive primer dimer bands. If these bands are present, we will need to purify the desired band from a traditional agarose gel.

The gels come pre-stained with EtBr or SYBR, and are encased in plastic. Buffer is already included, and the gels run rapidly (~ 10 min). The E-gel electrophoresis unit and a specific ladder are required.

1. Combine 16 µl diH20 and 4 µl PCR product.

2. Remove the comb from the gel.

3. Load into E gel wells. Load 20 µl diH20 into empty wells.

4. Load 10 µl low range ladder

5. Depending on the gel base, choose the proper program if available and begin electrophoresis.

6. Run for the specified amount of time.

7. Remove E gel and image in UV box.

16S rRNA gene products will be roughly 300-350 bp. ITS amplification products will be roughly 1000-1100 bp.

8. Record successful reactions on the E-gel layout spreadsheet.

3. DNA Quantification with the Qubit fluorometer (*Sample dependent about 90 minsfor 96 samples*)

a. Introduction

The Qubit system was designed to specifically quantify nucleic acids and proteins using small quantities of PCR product. The fluorescent probe (Qubit reagent) intercalates double stranded DNA (dsDNA) and fluoresces only after intercalation. Other methods of DNA quantification rely on UV-Vis spectroscopy to quantify nucleic acids; however they are much less specific as the dsDNA, RNA, and proteins absorb overlapping wavelengths. Since the fluorophore fluoresces only after intercalating dsDNA, the DNA concentrations assayed with the Qubit system are much more accurate than with other methods. See the Qubit 2.0 user manual and the Invitrogen website for more information[4].
(http://www.invitrogen.com/site/us/en/home/brands/Product-Brand/Qubit/qubit-fluorometer.html)

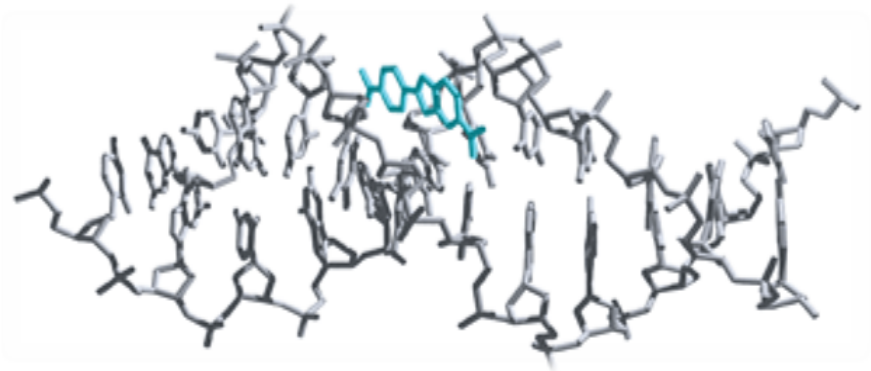


Figure 3. The fluorescent probe (blue) intercalates the dsDNA, allowing for both precise measurement of the dsDNA concentration of a sample. For more information, see the

Invitrogen website. (http://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/DNA-RNA-Purification-Analysis.html)

b. Materials

Qubit Fluorometer
Qubit dsDNA HS Buffer
Qubit reagent (protect from light)
Qubit Assay tubes
Standards 1 and 2, concentrations of 0ng and 10ng respectively
DNA extract or PCR Product

c. Protocol

Manufacturer's Diagram

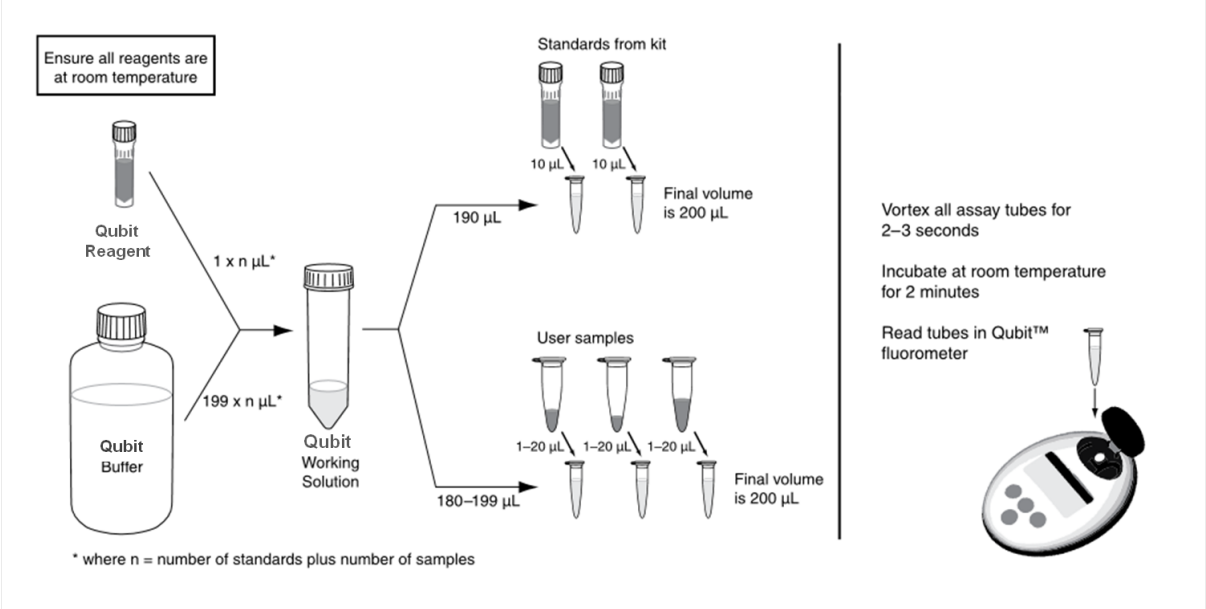


Figure 4. Manufacturer's diagram of the Qubit protocol. See the Qubit 2.0 for the high sensitivity dsDNA manual for more information[5].

1. Prepare working buffer:

    Qubit dsDNA HS Buffer: [Number of samples+3]*199µl = _____

    Qubit reagent (fluorophore): [Number of samples+3]*1µl = _____

    Note: The extra 3 samples allow for 2 standards and for pipetting error

2. Vortex the working buffer to mix
3. Label Qubit Assay tubes with sample ID
4. For each sample, add 2µl of PCR product to 198µl of working buffer to the appropriate tube
5. For each of the two standards, add 10µl of standard to 190µl of working buffer to the appropriate tube
6. Vortex each sample for 2-3 seconds to mix
7. Incubate for 2 minutes at room temperature

8. On the Qubit fluorometer, hit ***DNA***, then ***dsDNA High Sensitivity***, then ***YES***.

9. When directed, insert standard 1, close the lid, and hit ***Read***

10. Repeat step 9 for standard 2. This produces your two-point standard calibration.

11. Read each sample by inserting the tube into the fluorometer, closing the lid, and hitting ***Read Next Sample***

12. Use the spreadsheet *dna_quants.xlsx* to record the data.

D. Quality Check Libraries

1. Pool 240 ng DNA per sample into one collection tube (*one hour*)
See the column "Volume to pool" in the spreadsheet *dna_quants.xlsx* for the volume of each sample to add to the pool.

2. Gel Electrophoresis (prep 90 mins; loading and runtime 90 mins)
There may be extra bands in the PCR product caused by amplification of chloroplast DNA, primer dimers, or other undesired amplification. You can separate the desired band from the unwanted bands by traditional gel electrophoresis and gel purification.

1. Tape up gel tray or use rubber gasket to seal.

2. Place 4 combs into gel tray at every 3 notches. ***Do this before you pour the gel!

3. Make a 2% agarose gel

    a) Weigh out 8 g agarose

    b) Make 400 ml 1X TBE buffer (80 ml 5X TBE, 320 ml DI H_20)

    c) Microwave gel mixture to combine ingredients,

        **check every 30-40 sec, watch out for boiling over!!

        ** wear a full glove to protect your hand and forearm in case it does boil over

4. Cool gel solution on a stir plate (NO HEAT!) until you can touch the glass

5. Add 4 µl Gelstar (SYBR green based) per 100 ml gel to the gel (16 µl) right before you pour.

6. Pour gel slowly from one corner. Avoid introducing air bubbles as you pour.

7. Let gel cool for between 1 and 1.5 hours. Remove tape and place into gel rig.

8. Make 2500 ml 0.5X TBE running buffer (big rig). Make sure it's enough to cover gel.

9. Make up 50 ml of loading buffer (keep in refrigerator)

    a) 15 g sucrose

    b) 0.175 g Orange G

    c) bring up to 50 ml with DI H_2O

10. Load 4 ul orange g into each well of a 96 well plate.

11. Load 4 ul PCR product into each well of the same 96 well plate (DO NOT MIX!)

12. Remove combs gently from gel.
13. Load 5 ul ladder (AlphaQuant 4) into the first and last wells on each row.
14. Load all 8 ul PCR product/Orange G into the wells.

    Use a multichannel pipetter!

    **See accompanying spreadsheet for loading scheme

    **you may rinse tips in tank buffer and change them every few times.
15. Run gel for about 1.5 to 2 hrs at between 60-80 V.

3. QIAquick Gel Purification (2 hours)

1. Excise the DNA fragments from the agarose gel with a clean, sharp scalpel-
   - Minimize the size of the gel slice by removing extra agarose
     ** minimize light exposure and manipulation of gel as this can denature the DNA**
     *** ALWAYS wear safety glasses and keep the cover on the gel when looking on the light***
2. Weigh the gel slice in a colorless tube. Add 3 volumes of Buffer QG to 1 volume of gel.
   - E.g. a 100 mg gel slice would require 300 μL of Buffer QG. The maximum amount of gel slice per QIAquick column in 400 mg. For a gel slice > 400 mg use more than one QIAquick column.
3. Incubate at 50^{o}C for 10 minutes or until gel slice is completely dissolved. Can vortex to help dissolve gel mixture.
4. After gel slice is dissolved completely, check that the color of the mixture is yellow. If it is orange or violet, add 10 μL of 3 M sodium acetate, pH5 and mix. This should bring the mixture back to yellow.
5. Add 1 gel volume of isopropanol (or 200 proof EtOH) to the sample and mix.
6. Place a QIAquick spin column in a 2 mL collection tube
7. To bind DNA, apply the sample to the QIAquick column and centrifuge 1 minute. The maximum reservoir or the column is 800 μL. For samples greater than 800 μ just load and spin again.
8. Discard flow through and place column back in same collection tube.
9. Add 0.5 mL of buffer QG and centrifuge for 1 min.
10. To wash: add 0.75 mL buffer PE to QIAquick column and centrifuge 1 min.
11. Discard flow through and centrifuge for an additional 1 minute at 17,900g (13,000 rpm)
12. Place QIAquick column into a clean, labeled, 1.5 mL microcentrifuge tube
13. To elute DNA, add 30 μL of Buffer EB to the center of the QIAquick membrane and centrifuge for 1 minute. Take flow-through and spin through the column again. Discard column.
14. Freeze products.

4. Bioanalyzer (45 mins to set up chip; 45 mins for scanning)

a. Introduction

The Bioanalyzer (Agilent) is used to assess the quality of the pooled DNA before it is sent to the sequencing core. The Bioanalyzer uses microfluidics technology to carry out gel electrophoresis on a very small scale. A gel-dye mix is prepared and spread into the wells of the chip during the chip priming step. Marker, the ladder, and the samples are loaded and the chip is vortexed briefly. During the run, the DNA fragments migrate and are compared to the migration of the ladder, resulting in a precise calculation of DNA fragment size and abundance. The Bioanalyzer works with RNA as well, and is useful for determining the quality of RNA. See the DNA assay protocol and the Agilent website for more information about the applications and troubleshooting guides[6].

http://www.genomics.agilent.com/en/Bioanalyzer-Instruments-Software/2100-Bioanalyzer/?cid=AG-PT-106&tabId=AG-PR-1001

b. Agilent High Sensitivity DNA Assay Protocol

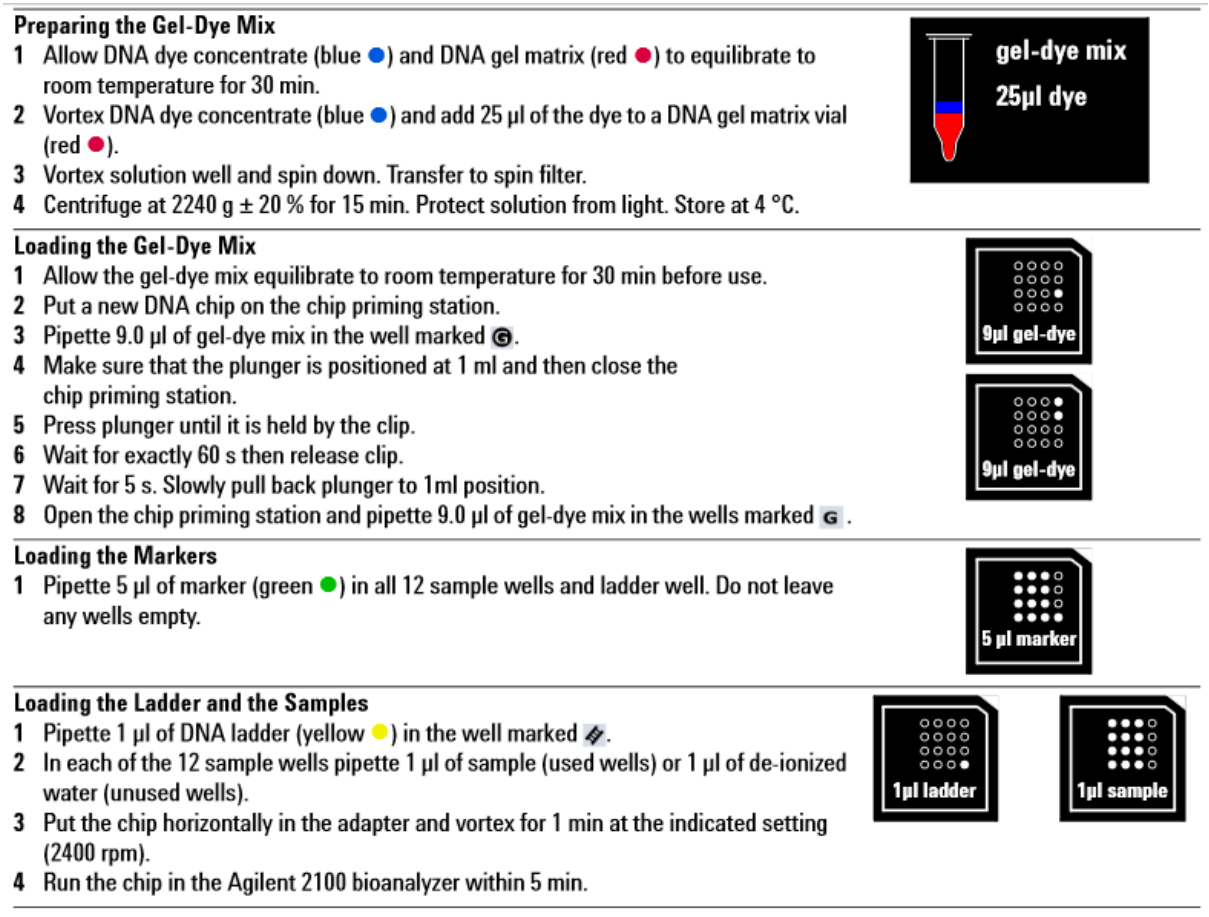

Preparing the Gel-Dye Mix
1 Allow DNA dye concentrate (blue) and DNA gel matrix (red) to equilibrate to room temperature for 30 min.
2 Vortex DNA dye concentrate (blue) and add 25 µl of the dye to a DNA gel matrix vial (red).
3 Vortex solution well and spin down. Transfer to spin filter.
4 Centrifuge at 2240 g ± 20 % for 15 min. Protect solution from light. Store at 4 °C.
gel-dye mix
25µl dye
Loading the Gel-Dye Mix
1 Allow the gel-dye mix equilibrate to room temperature for 30 min before use.
2 Put a new DNA chip on the chip priming station.
3 Pipette 9.0 µl of gel-dye mix in the well marked G.
4 Make sure that the plunger is positioned at 1 ml and then close the chip priming station.
5 Press plunger until it is held by the clip.
6 Wait for exactly 60 s then release clip.
7 Wait for 5 s. Slowly pull back plunger to 1ml position.
8 Open the chip priming station and pipette 9.0 µl of gel-dye mix in the wells marked G.
9µl gel-dye
9µl gel-dye
Loading the Markers
1 Pipette 5 µl of marker (green) in all 12 sample wells and ladder well. Do not leave any wells empty.
5 µl marker
Loading the Ladder and the Samples
1 Pipette 1 µl of DNA ladder (yellow) in the well marked.
2 In each of the 12 sample wells pipette 1 µl of sample (used wells) or 1 µl of de-ionized water (unused wells).
3 Put the chip horizontally in the adapter and vortex for 1 min at the indicated setting (2400 rpm).
4 Run the chip in the Agilent 2100 bioanalyzer within 5 min.
1µl ladder
1µl sample

Figure 5. Agilent Bioanalyzer Protocol Overview. See the Quick Start guide for more information[7].

Preparing the Gel Dye Mix

1. Allow High Sensitivity DNA dye concentrate (blue●) and High Sensitivity DNA gel matrix (red●) to equilibrate to room temperature for 30 min.

2. Add 15 µl of High Sensitivity DNA dye concentrate (blue●) to a High Sensitivity DNA gel matrix vial (red●).
3. Vortex solution well and spin down. Transfer to spin filter.
4. Centrifuge at 2240 g +/- 20% for 10 min. Protect solution from light. Store at 4 °C.

Loading the Gel-Dye Mix
1. Allow the gel-dye mix to equilibrate at room temperature for 30 min before use.
2. Put a new High Sensitivity DNA chip on the chip priming station.
3. Pipette 9.0 µl of gel-dye mix in the well marked (G)
4. Make sure that the plunger is positioned at 1 ml and then close the chip priming station.
5. Press plunger until it is held by the clip.
6. Wait for exactly 60 s then release clip.
7. Wait for 5 s, then slowly pull back the plunger to the 1 ml position.
8. Open the chip priming station and pipette 9.0 µl of gel-dye mix in the wells marked (G).

Loading the Marker
1. Pipette 5 µl of marker (green●) in all sample and ladder wells. Do not leave any wells empty.

Loading the Ladder and the Samples
1. Pipette 1 µl of High Sensitivity DNA ladder (yellow●) in the well marked .
2. In each of the 11 sample wells pipette 1 µl of sample (used wells) or 1 µl of marker (unused wells).
3. Put the chip horizontally in the adapter and vortex for 1 min at the indicated setting (2400 rpm).
4. Run the chip in the Agilent 2100 Bioanalyzer within 5 min.

c. Interpreting Bioanalyzer Results

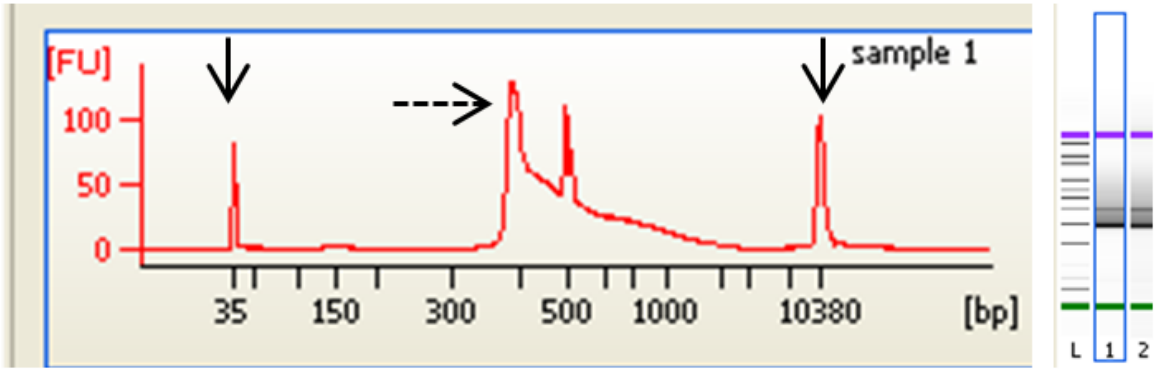


Figure 6. An example tracing from the Bioanalyzer DNA assay containing a high quality barcoded 16S V4 amplicons.

The peaks at the 35 and 10,380 bp are the marker peaks (black solid arrows). The peak around 300 bp is the peak of interest, and represents the approximately 360 bp V4 barcoded amplicon. Sometimes extraneous peaks are present, like the peak around 500 bp. A small bump in the tracing is seen around 150 bp, which indicates there is a small amount of primer still left in the sample; however, this peak is insignificant compared to the strong peak corresponding to the barcoded amplicon. The gel to the right corresponds to the peaks, with the most intense sample band slightly less than 400 bp.

Assessments:

1) Content Assessments

For more detailed content assessments see Week 1through Week 5 Assessment folders for the Environmental Genomics course on the Wiki or on your flash drive.

List of assessment questions/activities:

- Describe why the 16S rRNA gene is a good phylogenetic marker
- Describe the benefits of replication in designing a 16S rRNA gene study.
- Design a study to compare microbial community structure in your system of choice (from environmental sample to data analysis).
- Draw the design of barcoded Illumina 16S rRNA gene targeted primers (Forward and Reverse). Label each section of the primer and describe its function in library construction.
- Summarize the steps involved in preparing Illumina barcoded 16S rRNA gene libraries as described in the Caporoso et al paper.
- Discuss biases associated sample preparation (from collection through library preparation) that might result in biases in microbial community structure.
- Describe the steps of Illumina sequencing.

2) Molecular Techniques Post Course Student Attitudes Survey

This survey can be given at the end of the modules. Also it can be found in the "Post-course survey folder" on the Wiki and the flash-drive

1. Professional information (please circle all relevant descriptions)

   a. Elementary School Teacher
   b. Middle school teacher
   c. High School Teacher
   d. College faculty/staff
   e. Student/Graduate Student
   f. Writer/Artist/Creator
   g. Other_____

2. Please indicate your primary academic disciplinary area below.

3. Which of the following best describes your previous experience with scientific research?
   a. this is my first research experience
   b. I have had some limited research experience prior to this course (less than one year)
   c. I have had 1-2 years of research experience
   d. I have more than 2 years of research experience
   e. other_____

4. Reason for taking Molecular Techniques

a. Couldn't fit any other class in your schedule
b. Wanted to learn about and apply cutting edge molecular technologies
c. General Interest
d. Other_____

5. Gender
a. Female
b. Male
c. Other_____
d. prefer not to answer

6. Molecular techniques Assessment *(circle your response)*

	Very unsatisfied	**Unsatisfied**	**Neutral**	**Satisfied**	**Very Satisfied**	**Not Applicable**
Overall Experience	1	2	3	4	5	
Laboratory experience	1	2	3	4	5	
Bioinformatics experience	1	2	3	4	5	
Biostatistical Experience	1	2	3	4	5	
Scientific Writing experience	1	2	3	4	5	
Quizzes	1	2	3	4	5	
Assignments	1	2	3	4	5	
Professor	1	2	3	4	5	
Handouts	1	2	3	4	5	
Discussions	1	2	3	4	5	

7. **Pre/Post assessment:** Please assess each of the following in terms of how you felt ***BEFORE*** attending Molecular techniques and how you feel ***NOW***.

7A	**Very unlikely**	**Somewhat unlikely**	**Neutral**	**Somewhat likely**	**Very likely**
Likelihood of using next generation sequencing technologies in research – BEFORE.	1	2	3	4	5
Likelihood of using next generation sequencing technologies in research – NOW.	1	2	3	4	5

7B	**Very low**	**Somewhat low**	**Neutral**	**Somewhat high**	**Very high**
Knowledge of bioinformatics. – BEFORE.	1	2	3	4	5
Knowledge of bioinformatics – NOW.	1	2	3	4	5

7C	**Very low**	**Somewhat low**	**Neutral**	**Somewhat high**	**Very high**
Knowledge of biostatistical approaches – BEFORE.	1	2	3	4	5

Knowledge of biostatistical approaches – NOW.	1	2	3	4	5

7D	**Very low**	**Somewhat low**	**Neutral**	**Somewhat high**	**Very high**
Knowledge of unix/linux operating systems – BEFORE.	1	2	3	4	5
Knowledge of unix/linux operating systems – NOW.	1	2	3	4	5

7E	**Very low**	**Somewhat low**	**Neutral**	**Somewhat high**	**Very high**
Comfort in executing command line based programs – BEFORE.	1	2	3	4	5
Comfort in executing command line based programs – NOW.	1	2	3	4	5

Open-Ended Questions

8. What were the strengths of the Molecular Techniques course? What did you find most useful or enjoyable?
9. Which parts of the molecular techniques course were the least useful or enjoyable?
10. How likely are you to recommend this course to a friend or colleague?

Very unlikely	Somewhat unlikely	Neutral	Somewhat likely	Very likely
1	2	3	4	5

11. Do you have any other comments or suggestions for improving Molecular techniques?

12. What did you learn in Molecular techniques?

13. How did this course challenge you?

Time line of module

We will be performing all of the protocols described in this module over the 2.5 day workshop. It should be noted that this module can be spread out over the first five weeks of a semester. Please see the section on "applications in the classroom" for a detailed outline of how to do this. Also on your flash-drives and on the Wiki there will be links to all of the materials (week by week) for setting this up as a semester long class.

Discussion Topics for class

- Experimental design
  - Why is replication important in biological studies?
  - What are some simple statistical techniques that one can use if they don't replicate sampling?
  - What is the difference between biological and technical replicates?

- Describe how replication enables one to make more inference about the potential differences in bacterial composition between two samples.
- Discuss how the secondary structure of the gene is relevant to its evolution and function.
- Discuss how Carl Woese discovered the third domain of life
- Describe the utility of multiplexing samples using this approach
- Discuss biases associated with each section of the modules, especially PCR biases.
- Discuss the importance of quality checking DNA to be sequenced
- Discuss Illumina sequencing chemistry

Applications in the classroom

Environmental Genomics Research Course Syllabus

Meeting Time: TBA two, three hour sessions

Goals:

- Students will learn and apply, hands-on novel molecular techniques to study microbial communities in the context of an environmental or health related problem.
- Students will learn how to generate microbial DNA libraries for high-throughput sequencing, and use appropriate informatics and statistical workflows to analyze sequence data they generate to answer biologically meaningful research questions.

Required Text: Samuelsson, Tore. Genomics and Bioinformatics: An Introduction to Programming Tools for Life Scientists. Cambridge University Press. 2012.

Course Objectives:

- Students will design and execute a project where they will investigate the microbial community profiles from samples of environmental or health related significance. For the first half of the semester students will perform and be able to describe the biochemistry behind nucleic acid extraction, quantification, and 16S rRNA gene PCR to generate libraries for sequencing. *Example Project: Temporal Dynamics of Microbial Community Structure of stormwater collected downstream of a combined sewer overflow*
- Students will explain the biochemistry behind the most recent high throughput sequencing technologies and perform cost benefit analysis of utilizing different sequencing applications.
- Students will apply unix and perl based bioinformatics tools to perform computational analysis on various types of genomics projects, including the sequence data they generate from their semester long research project.

- Students will prepare a scientific manuscript in which they synthesize the current literature relevant to their research problem, describe their methodology, and present and discuss their research findings.
- Each student will develop a poster describing the technology behind a molecular technique of their choice.
- Through discussion of current literature in the field, students will develop plans to troubleshoot experimental and bioinformatics problems that they may encounter.
- Exposure to this type of research will also catalyze advanced undergraduate training in the integration of basic biological concepts, cutting edge, modern sequencing technologies and bioinformatics with the multi and cross disciplinary approaches.

Assessment:
25% quizzes- Drop your lowest. Pre and Post quizzes for each topic
10% assignments
10% Poster of a molecular technique
10% presentation of a bioinformatics exercise from Samuelsson.
10% Final Presentation
35% manuscript

Grading will be as follows:

A = Your work must be of the quality and completeness that could be published as part of a scientific manuscript. To earn an A you must also demonstrate a high level of independence. (i.e. when something goes wrong you need to formulate a plan to figure things out) Of course I will help discuss with you a proper way to proceed but you need to initiate plans of troubleshooting. You effectively organize your time, carefully plan experiments and critically assess your work. Additionally you clearly and professionally can communicate your project (written and oral).

B = Your work is of good quality, but is not sufficiently complete to warrant an A.

C = Only basic requirements listed above are met.

D/F = Student does not meet the course requirements and will most likely be recommended to drop the class unless they can demonstrate/formulate a plan to get themselves back on track.

Academic Integrity: I expect that each of you will put forth your best honest effort in this class. I also expect that you are responsible for upholding the standards of academic honesty and integrity as described by Juniata College. Please familiarize yourself with these policies, which can be found at:
http://www.juniata.edu/services/pathfinder/Academic_Honesty/standards.html

Students with Disabilities: If you need special accommodations with respect to disabilities please contact the academic counselor in Academic Support Services who will help you with this process. More information can be found at:
http://www.juniata.edu/services/catalog/section.html?s1=appr&s2=accommodations**Course**

Withdrawal: After the drop/add period has expired, you may withdraw from this class. Your last chance to do this is the last day of classes. You will need my signature and the signature of your advisor(s).

This module took approximately 5 weeks to teach in the classroom. From nucleic acid extraction through sending out the libraries for sequencing. Below you can find the objectives for each week. All of the course materials can be found in the Environmental Genomics Research folder on your flashdrive. Course materials are organized by week. Keep in mind this was my first year teaching the course, so approach cautiously.

Week 1: rRNA structure and function and nucleic acid extraction
Outline of Objectives

- Be able to define the structure and function of the rRNA operon
- Utilize the 16S rRNA gene technology to describe microbial community structure in a sample.
- Describe the general steps involved in DNA/RNA extraction.
- Be able to troubleshoot problems associated with nucleic acid extraction.
- Perform DNA/RNA purifications and describe the underlying biochemistry behind each step
- Define sources of variation in each step of our experimental sampling and design methods to measure this variation.
- Understand the difference between biological and technical replication and provide examples of each in the context of our experimental design.
- Describe the sources of error in a 16S rRNA gene study.
- Design a replication strategy for our 16S rRNA gene study.
- Describe some statistical methods that we can use if replication is not feasible.

Week 2: Illumina tag PCR
Outline of Objectives

- Describe biases associated with PCR and how they might effect microbial community analysis.
- Understand how the Illumina itag PCR works. (i.e. be able to draw the forward and reverse constructs and know the function of each portion of the construct). Also be able to draw the first three cycles of PCR
- Evaluate potential strategies for overcoming these different PCR biases.
- Define different ways to measure DNA concentration
- Perform itag PCR on our environmental samples
- Discuss the various biases associated with different regions of the 16S rRNA gene
- Explain how the secondary structure of the gene is relevant to its evolution and function.

- Introduce various technologies that each group with be making poster for (DNA/RNA co-extraction, itag PCR, the Qubit & Bioanalyzer (group of three), E-gels, q-PCR, Illumina sequencing.
- Utilize a bufferless gel system to run a gel to check PCR products.
- Troubleshoot PCR reactions that didn't work.

Weeks 3 and 4: Troubleshooting and PCR product purification
Outline of Objectives

- Continue troubleshooting negative PCR reactions
- Utilize SPRI bead technology and/or gel extraction to clean up pooled PCR products
- Explain the biochemistry behind SPRI bead technology
- Learn how to evaluate the quantity and quality of the prepared libraries
    - Using gel electrophoresis
    - Bioanalyzer
    - Qubit
- Describe how to perform a literature review for the manuscript
- Describe how Carl Woese discovered the third domain of life. (consider moving to week 1 or 2)
- Discuss why fecal bacteria indicators do not always correlate with pathogens (**project specific objective)
- Describe limitations of pathogen detection (**project specific objective)
- Evaluate the advantages and disadvantages of shotgun metagenomics
-
- **Week 5: Sequencing Technologies**
    - **Outline of Objectives**
-
- Watch the following Broad Institute Bootcamp videos http://www.broadinstitute.org/scientific-community/science/platforms/genome-sequencing/broadillumina-genome-analyzer-boot-camp
- Be able to describe the biochemistry behind the Sanger, 454 and Illumina sequencing technologies.
- Analyze the output of the 454 and Illumina sequencing platforms.
- Explain the limitations and advantages of these sequencing platforms.
- Understand the Illumina sequencing technology in the context of our application. Make them draw a couple of cycles of paired end Illumina sequencing. (I also make them dig through the patents to find relevant information)

References and Suggested Reading

High-throughput sequencing of 16S rRNA

1. Earth Microbiome Project. http://www.earthmicrobiome.org/
2. Caporaso JG, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7:335–336.
3. Caporaso JG, et al. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U.S.A. 108 Suppl 1:4516–4522.
4. Kuczynski J, et al. 2011. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Bioinformatics Chapter 10:Unit 10.7.
5. Soergel DAW, Dey N, Knight R, Brenner SE. 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME Journal 6: 1440–1444.

Sample Design and Replication

6. Lennon JT. 2011. Replication, lies and lesser-known truths regarding experimental design in environmental microbiology. Environmental Microbiology 13:1383–1386.
7. Prosser JI. 2010. Replicate or lie. Environmental Microbiology 12:1806–1810.

ITS region papers

8. McGuire KL, et al. 2013. Digging the New York City skyline: Soil fungal communities in green roofs and city parks. PLoS ONE 8:e58020.
9. Kõljalg U, Larsson K-H, Abarenkov K, Nilsson RH, Alexander IJ, et al. 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytol 166: 1063–1068.
10. Peay K.G., Kennedy P.G., Bruns T.D. 2008. "Fungal community ecology: a hybrid beast with a molecular master". BioScience 58: 799–810.

Background on rRNA operon and discovery of archaea

11. Balch WE, Magrum LJ, Fox GE, Wolfe RS, Woese CR. An ancient divergence among the bacteria. J Mol Evol. 1977 Aug 5;9(4):305-11.
12. Fox GE, Magrum LJ, Balch WE, Wolfe RS, Woese CR. Classification of methanogenic bacteria by 16S ribosomal RNA characterization. Proc Natl Acad Sci U S A. 1977 Oct;74(10):4537-4541.
13. Woese CR, Fox GE**.** Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 1977 Nov;74(11):5088-90.

Databases

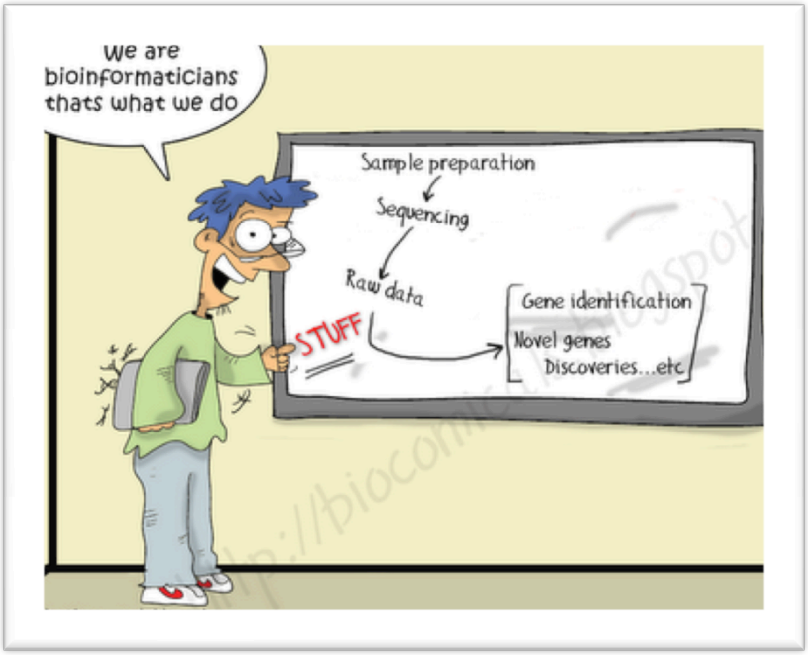
UNITE database for fungal identification (http://unite.ut.ee/)

Ribosomal Database Project (http://rdp.cme.msu.edu/)

Greengenes Database: (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi)

ARB-Silva Database (http://www.arb-silva.de/)

MODULE 2: ANALYSIS OF SEQUENCE DATA



This module focuses on de-convoluting the 'stuff'.

Background

The Illumina MiSeq platform allows cost-effective deep sequencing by offering multiplexing and acceptable read lengths with a short turnaround time. The open source Quantitative Insights into Microbial Ecology (QIIME) pipeline, which runs in Linux environments, handles reads from a variety of sequencing platforms and can be used to process the raw reads, generate an OTU table, perform diversity analyses, and compute statistics. This module aims to introduce the shell and useful commands in Linux, and to provide a short overview of the capabilities of QIIME for microbial community analysis.

Module goals

The goal of this module is to become familiar with the Linux file structure and command line, as well as to use the QIIME to analyze 16S rRNA and ITS data from sequences

Participants will also learn how to properly utilize multivariate statistics to help answer their biological questions.

Vision and Change core competencies addressed

- Ability to apply the process of science by developing hypotheses and designing bioinformatics/biostatistical workflows relevant to understanding microbial communities in their natural environments.

- Ability to use quantitative reasoning by:
    - developing and Interpreting alpha and beta diversity graphs
    - Applying statistical methods to understand the relationship between microorganisms and their environment
    - Using models of microbial diversity to explain observed data
    - Using bioinformatics and biostatistical tools for analyzing large sequence data sets

GCAT-SEEK sequencing requirements

See description in Module 1.

Computer/program requirements for data analysis

- Microsoft Excel or similar program
- Oracle VirtualBox (free, see appendix)
- QIIME 1.7 (free)
- Server or cluster with multiple cores (You will have remote access to both the GCAT-SEEK and HHMI clusters.

Protocols

A. Unix/Linux Tutorial

Linux is an open-source Unix-like operating system. It allows the user considerable flexibility and control over the computer by command line interaction. Many bioinformatics

pipelines are built for Unix/Linux environment; therefore it is a good idea to become familiar with Linux basics before beginning bioinformatics.

Every desktop computer uses an **operating system**. The most popular operating systems in use today are Windows, Mac OS, and UNIX.

Linux is an operating system very much like unix, and it has become very popular over the last several years. Operating systems are computer programs. An operating system is the first piece of software that the computer executes when you turn the machine on. The operating system loads itself into memory and begins managing the resources available on the computer. It then provides those resources to other applications that the user wants to execute.

The shell- The shell acts as an interface between the user and the kernel. When a user logs in, the login program checks the username and password, and then starts another program called the shell. The shell is a command line interpreter (CLI). It interprets the commands the user types in and arranges for them to be carried out. The commands are themselves programs: when they terminate, the shell gives the user another prompt ($ on our systems).

Filename Completion - By typing part of the name of a command, filename or directory and pressing the [**Tab**] key, the shell will complete the rest of the name automatically. If the shell finds more than one name beginning with those letters you have typed, it will pause, prompting you to type a few more letters before pressing the tab key again.

History - The shell keeps a list of the commands you have typed in. If you need to repeat a command, use the cursor keys to scroll up and down the list or type history for a list of previous commands.

Files and Processes
Everything in UNIX is either a file or a process.
A process is an executing program identified by a unique process identifier. A file is a collection of data. They are created by users using text editors, running compilers etc.
Examples of files:

- a document (report, essay etc.)
- the text of a program written in some high-level programming language
- instructions comprehensible directly to the machine and incomprehensible to a casual user, for example, a collection of binary digits (an executable or binary file);
- a directory, containing information about its contents, which may be a mixture of other directories (subdirectories) and ordinary files.

It is not required to have a Linux operating system to use QIIME. While a native installation is possible, many people use the version of QIIME distributed inside of a VirtualBox. The VirtualBox is a free product from Oracle, the company that maintains Java. VirtualBox allows a segment of your computer to be virtualized and run a different operating system. The QIIME developers package QIIME and all of it its dependencies inside a virtual disk image (VDI), which functions as a virtual hard drive with pre-installed components. We will use the VirtualBox here. Detailed instructions for installing the VirtualBox and creating a virtual machine with the QIIME VDI are in the appendix.

Once you have the Linux operating system running, we can explore a few features of the Ubuntu distribution. The graphical user interface (GUI, "gooey") looks like the typical desktop of a PC. The dashboard on the left shows all of the frequently used or the running applications. The System Monitor is the task manager of Ubuntu, and you can see how many resources (RAM, CPU, etc.) are being used here. You can navigate using the graphical file structure by clicking on the directories (folders) on the desktop or by using the terminal.

In this example, we've navigated to the Desktop of the qiime user from the root directory. The directories Before_you_start and Shared_Folder can also be seen in the terminal.

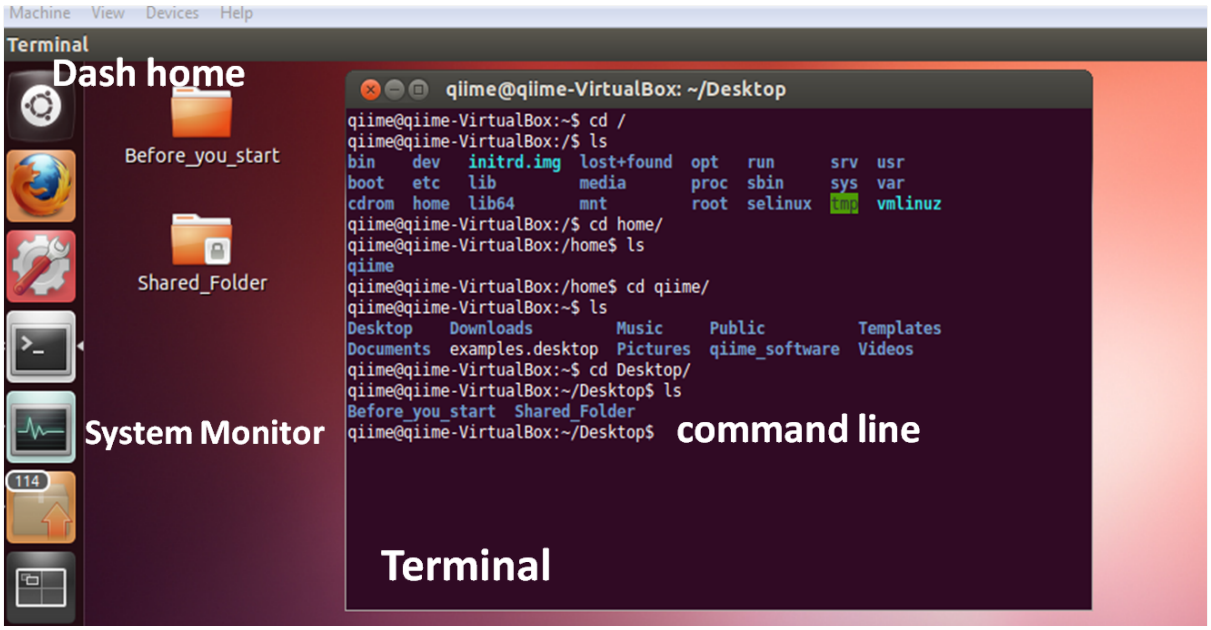


Figure 7. Several components of the Ubuntu graphical user interface (GUI). The dash home contains all of the programs installed. The system monitor allows you to check resource use. The terminal is the means by which you can interact with the computer through the command line.

Understanding the file structure and knowing how to use some basic Linux commands are essential for using QIIME effectively. Below is a simplified version of the file structure we see in our distribution of Linux.

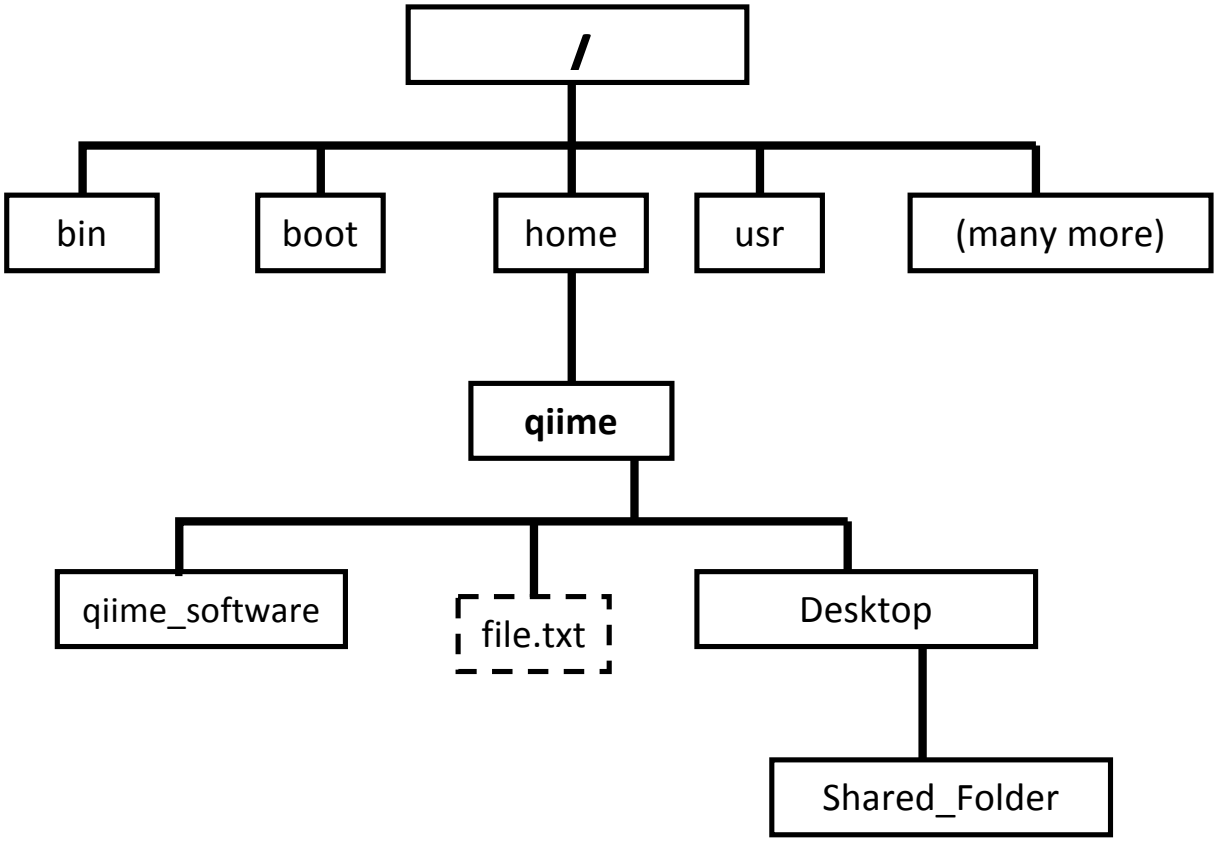


Figure 8. A simplified Linux file structure. Directories are in solid boxes, while files are denoted by dashed boxes. The root directory is shown as /. The subdirectories within qiime are qiime_software and Desktop.

Question What subdirectories are within root?

The file structure is important when we use the command line, since we need to tell the shell where to find certain files, or where to output the results, for example. The full path to qiime is /home/qiime/, where the first forward slash indicates root.

Question What is the full path for Shared_Folder?

The shell commands we use most often have to do with navigation within the file structure and file management.

1. Open the terminal.
2. Type pwd into the terminal and press enter to print the working directory. This will allow you to see the current directory.

```
qiime@qiime-VirtualBox:~$ pwd
/home/qiime
qiime@qiime-VirtualBox:~$
```

3. Use ls (list) to list the contents of your home directory. Which files and directories do you see?

```
qiime@qiime-VirtualBox:~$ ls
```

4. Type ls –a to see the hidden files in your home directory. Note there is a space between ls and the dash. What distinguishes a hidden file from a visible file?

5. Navigate to the Desktop by using cd (change directory). Don't forget to practice using tab to have the shell fill in as much information as possible.

```
qiime@qiime-VirtualBox:~$ cd Desktop/
qiime@qiime-VirtualBox:~/Desktop$
```

6. Create a directory using mkdir within Desktop. Check to see if it worked by using ls.

```
qiime@qiime-VirtualBox:~$ cd Desktop/
qiime@qiime-VirtualBox:~/Desktop$ mkdir dir_1
qiime@qiime-VirtualBox:~/Desktop$ ls
Before_you_start  dir_1  Shared_Folder
qiime@qiime-VirtualBox:~/Desktop$
```

7. Make two subdirectories called test_1 and test_2 within dir_1. Write down the commands you used.

8. Return to the home directory using cd .., which tells the shell to go up one directory. Check to make sure you are there by using pwd.

```
qiime@qiime-VirtualBox:~/Desktop/dir_1$ cd ..
qiime@qiime-VirtualBox:~/Desktop$ cd ..
qiime@qiime-VirtualBox:~$ pwd
/home/qiime
qiime@qiime-VirtualBox:~$
```

9. View file.txt with the text viewer less. What does the file contain? After you are finished, hit q to exit less.

```
qiime@qiime-VirtualBox:~$ less file.txt
qiime@qiime-VirtualBox:~$
```

10. Let's copy this very interesting file to test_1. Use the command cp and provide the file to copy and the path. Don't forget to use tab to autocomplete. How do you check to see if the command worked correctly?

```
qiime@qiime-VirtualBox:~$ cp file.txt Desktop/dir_1/test_1/
qiime@qiime-VirtualBox:~$
```

11. Move file.txt from test_1 to test_2 using the mv command. As with cp, it takes the file name to be moved and the output path as arguments. Write down the commands. One method of doing this is below.

```
qiime@qiime-VirtualBox:~$ cd Desktop/dir_1/
qiime@qiime-VirtualBox:~/Desktop/dir_1$ mv test_1/file.txt test_2/
qiime@qiime-VirtualBox:~/Desktop/dir_1$ cd test_2/
qiime@qiime-VirtualBox:~/Desktop/dir_1/test_2$ ls
file.txt
qiime@qiime-VirtualBox:~/Desktop/dir_1/test_2$ cd ..
qiime@qiime-VirtualBox:~/Desktop/dir_1$ cd test_1
qiime@qiime-VirtualBox:~/Desktop/dir_1/test_1$ ls
qiime@qiime-VirtualBox:~/Desktop/dir_1/test_1$
```

After the mv command, you can check with ls to verify that file.txt is in test_2 but no longer in test_1.

12. Remove file.txt and the directories you created in this tutorial with the rm command. The –r option tells rm to remove recursively, which allows you to remove all files and directories within a directory.

```
qiime@qiime-VirtualBox:~/Desktop/dir_1/test_1$ cd ..
qiime@qiime-VirtualBox:~/Desktop/dir_1$ cd test_2
qiime@qiime-VirtualBox:~/Desktop/dir_1/test_2$ rm file.txt
qiime@qiime-VirtualBox:~/Desktop/dir_1/test_2$ cd ..
qiime@qiime-VirtualBox:~/Desktop/dir_1$ cd ..
qiime@qiime-VirtualBox:~/Desktop$ rm -r dir_1
qiime@qiime-VirtualBox:~/Desktop$ ls
Before_you_start  Shared_Folder
qiime@qiime-VirtualBox:~/Desktop$
```

It is sometimes challenging to adapt to interacting with the command line rather than a GUI. There are a number of resources online, both in help forums like StackOverflow and forums for Linux enthusiasts. Check out the GNU documentation too. As you use QIIME, you will get lots of practice moving around the file structure and providing absolute paths. To reduce typing errors, use tab as much as possible. If the shell doesn't understand what you are pointing to, then the QIIME scripts will not either.

For Reference

Change directory	cd
Go up one directory	cd ..
Print working directory	pwd

List the contents of directory	ls
Remove a file (a directory)	rm <file_name> (rm –R <dir_name>)
Copy a file (a directory)	cp <file_name> <new_location> (cp –R <file_name> <new_location>)
Move a file	mv <file_name> <new_location>
Rename a file	mv <file_name> <new_file_name>
Concatenate files	cat <file_1> <file_2> > <combined_file>
Write to a file	> file_name (ls > list.txt the contents of the directory will be listed in a text file called list.)
Superuser	sudo <command> (password is qiime)

B. Thinking about your biological question(s)

Take a moment to think about the questions you are investigating. Are you interested in how members of the microbial community are varying with environmental parameters? With a specific treatment or disease state? Are you interested in the presence / absence, relative abundance, or both? Are you interested in rare members of the microbial community? Do you have any specific hypotheses? Take a moment to write these biological questions down in the space below. This will help you understand what statistical approaches you may want to use.

C. INTRODUCTION TO QIIME

High throughput sequencing generates hundreds of thousands of reads per sample; therefore bioinformatics tools are necessary to translate the raw reads into meaningful biological information. Bioinformatics pipelines like QIIME enable microbial community analysis, in that they support the processing and analysis of large datasets containing millions of reads. QIIME uses reference databases to assign taxonomic information to a string of nucleotides representing the 16S sequences of most of the microbes in a sample. The definition of species in microbiology is somewhat complicated, as the traditional rules for defining species do not apply to microbes. In general, if two organisms have 97% 16S rRNA sequence identity, they are considered to be of the same species, or the same **operational taxonomic unit** (OTU). QIIME supports several strategies for generating an OTU table, or a table of each OTU present in the entire dataset and its absolute abundance in each of the samples, which we will explore in the following tutorial.

The dataset included in this tutorial is from a longitudinal study of an alluvial stream microbial community during and after a major storm event. The dynamics of the microbial community of a stream fed primarily by runoff and stormwater are not well understood in the context of major storm events, especially when combined and sanitary sewer overflows (CSO, SSO) feed into the stream. The samples (n = 47) were sequenced on the Illumina MiSeq platform and were demultiplexed (i.e. barcodes were matched to sample IDs and subsequently removed) on the sequencer, which requires us to modify the typical QIIME workflow to get around the split libraries step.

QIIME (pronounced "chime") is an open source bioinformatics platform for microbial community analysis. QIIME runs in Unix/Linux and can be natively installed or used inside a VirtualBox. The QIIME pipeline allows you input files directly from the sequencing instrument, demultiplex barcoded samples, generate an OTU table, and perform downstream diversity and statistical analyses.The basic workflow starts with filtering the reads to ensure good quality data and splitting the libraries to match barcoded samples to the appropriate metadata. The next step is to generate an OTU table, which is often done by clustering the reads against the Greengenes reference database; a process which greatly speeds computation. After the OTU table is generated and assigned taxonomies, various downstream analyses can be implemented. QIIME offers support for a number of alpha and beta diversity metrics, data visualization, and multivariate statistics. Furthermore, files generated in QIIME can be used with several other software packages, including Microsoft Excel, PC-ORD, Primer E, and R.

D. Installing the QIIME VirtualBox Image

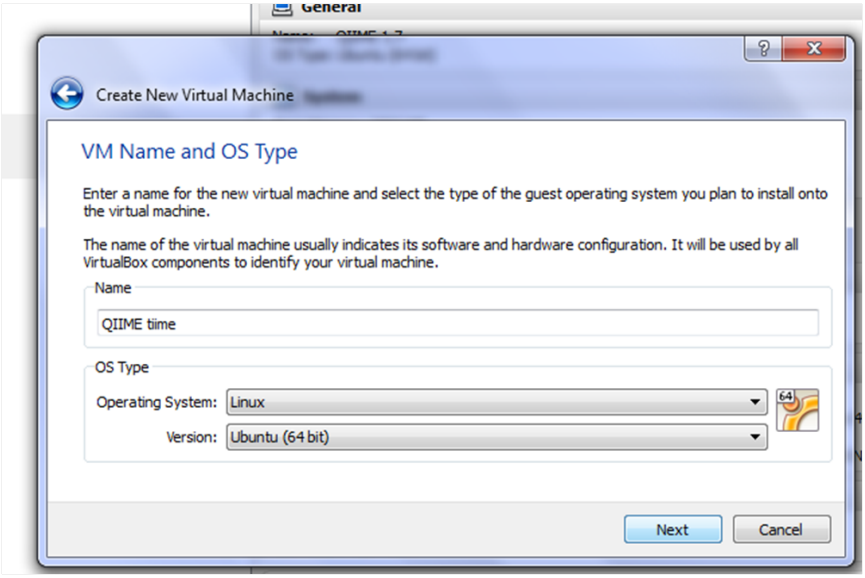
VirtualBox allows you to install QIIME, which runs in Linux, on a computer running a different operating system. Mac users can install QIIME for Macs, which is similar to a native installation of QIIME on a Linux operating system. This tutorial will cover installing VirtualBox (VB) for Windows users. You will need a **64-bit machine** and a tool to unzip files. 7zip works well, and is free and easy to install. http://www.7-zip.org/

There is a video tutorial available in which VirtualBox is installed on a Mac. The installation procedure should be similar to that as the Windows installation.
http://www.youtube.com/watch?v=1jYupkquaME

Download VB and the current QIIME virtual disk image (VDI). Follow the link below to install VirtualBox. To save time, the VDI is downloaded on the flash drive. It needs to be unzipped before installing. http://qiime.org/install/virtual_box.html

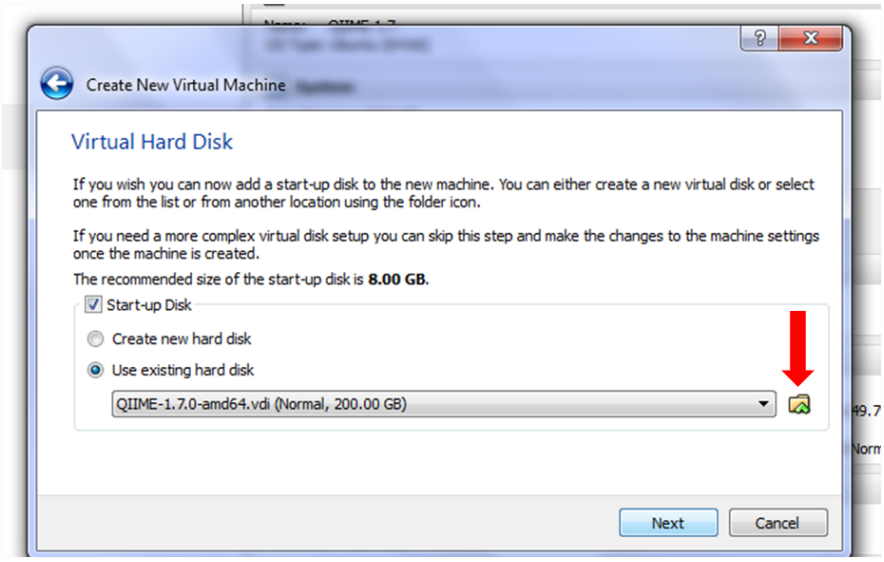
VirtualBox
Download VirtualBox
About
Screenshots
Downloads
Documentation
End-user docs
Technical docs
Contribute
Community
Here, you will find links to VirtualBox binaries and its source code.
VirtualBox binaries
By downloading, you agree to the terms and conditions of the respective license
VirtualBox platform packages. The binaries are released under the term
VirtualBox 4.2.12 for Windows hosts x86/amd64
VirtualBox 4.2.12 for OS X hosts x86/amd64
VirtualBox 4.2.12 for Linux hosts
VirtualBox 4.2.12 for Solaris hosts x86/amd64
VirtualBox 4.2.12 Oracle VM VirtualBox Extension Pack All supporte
Support for USB 2.0 devices, VirtualBox RDP and PXE boot for Intel cards.

Create a new virtual machine (VM)
Launch the VB manager and click "New" in the upper left hand corner. Give your new VM an informative name and change the operating system to **Linux** and the version to **Ubuntu (64-bit)**. Click "next" to continue the process.

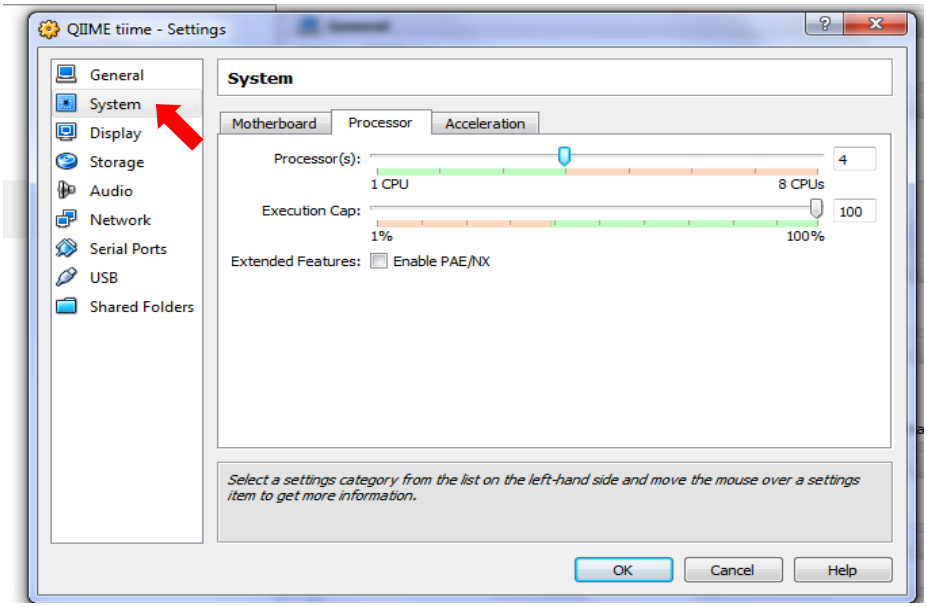

Create New Virtual Machine
VM Name and OS Type
Enter a name for the new virtual machine and select the type of the guest operating system you plan to install onto the virtual machine.
The name of the virtual machine usually indicates its software and hardware configuration. It will be used by all VirtualBox components to identify your virtual machine.
Name
QIIME time
OS Type
Operating System: Linux
Version: Ubuntu (64 bit)
Next
Cancel

Select the memory allocation to the VM.
Select at least 3 GB, but no more than indicated by the green region of the slider bar.

Load the QIIME VDI as the hard disk for the VM. Click on the folder (red arrow) to browse. Choose the unzipped VDI that you want to install.

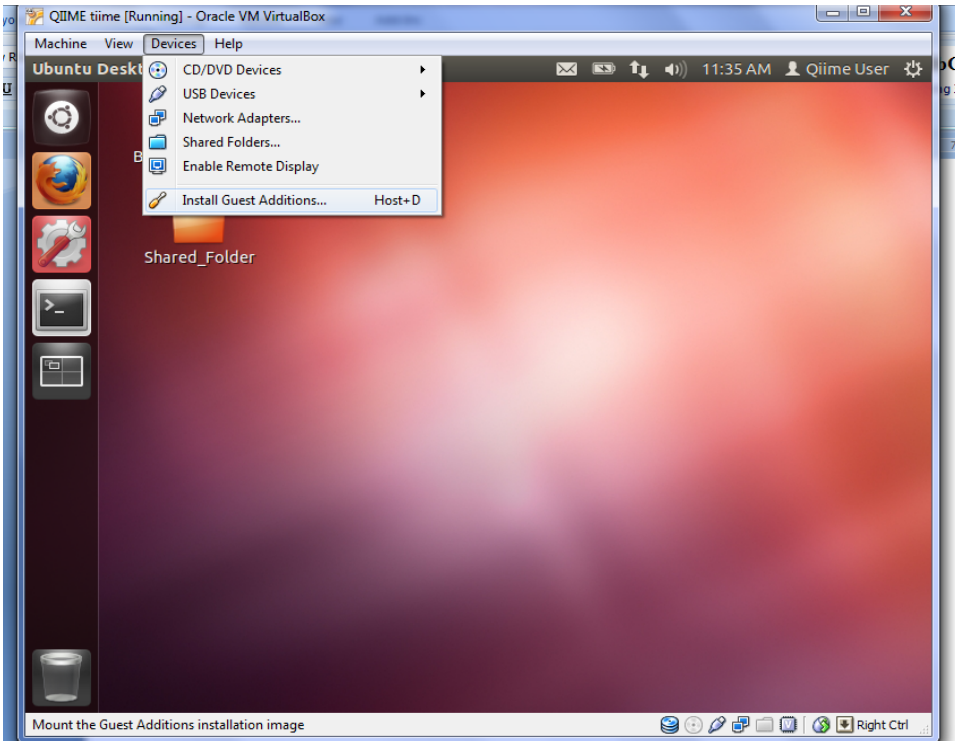

Create New Virtual Machine
Virtual Hard Disk
If you wish you can now add a start-up disk to the new machine. You can either create a new virtual disk or select one from the list or from another location using the folder icon.
If you need a more complex virtual disk setup you can skip this step and make the changes to the machine settings once the machine is created.
The recommended size of the start-up disk is 8.00 GB.
Start-up Disk
Create new hard disk
Use existing hard disk
QIIME-1.7.0-amd64.vdi (Normal, 200.00 GB)
Next
Cancel

Change the VM Settings
Click on the VM in the left panel so that it is highlighted. Click "Settings" in the upper left hand corner. Go to the "System" section and the "Processor" tab. Enable as many cores as you would like to allow the VM to use. This must be done before QIIME can run in parallel.



Install Guest Additions

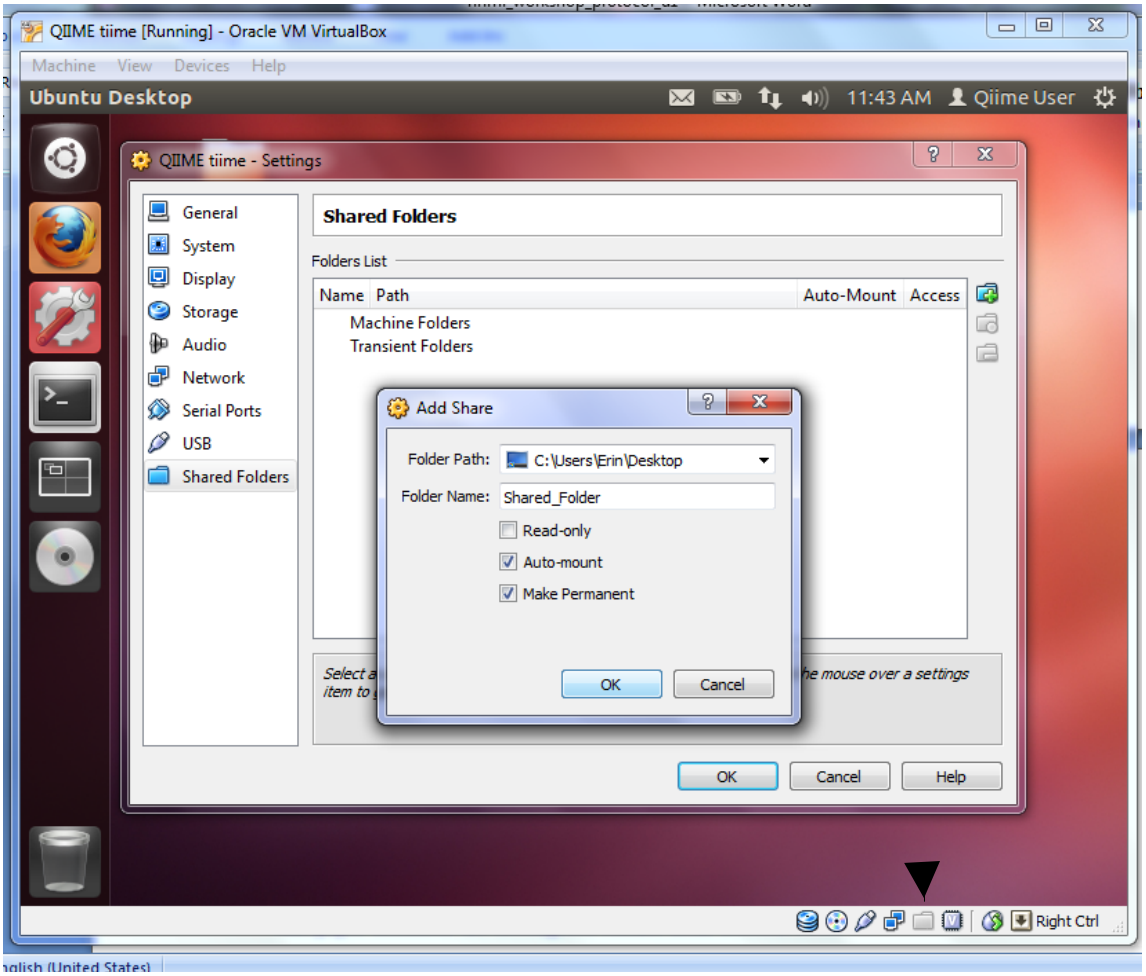
Click OK to get back to the VB Manager. Make sure your desired VM is highlighted and click "Start" or double click on the VM. The VM may take a few minutes to start up. If there are a few boot options to choose from, choose the first one by hitting enter. VB will prompt you for allowing the mouse and keyboard to be used inside VB. Hit okay to accept all of these prompts. Once the desktop is present, click "Devices" and "Install Guest Additions" at the bottom of the menu.



Click "Run" when the warning box pops up, and enter the password **qiime** when you are asked to authenticate the installation. You can monitor the process of installation from the command line. Installation may take a few minutes. When prompted, hit Enter to finish the installation.

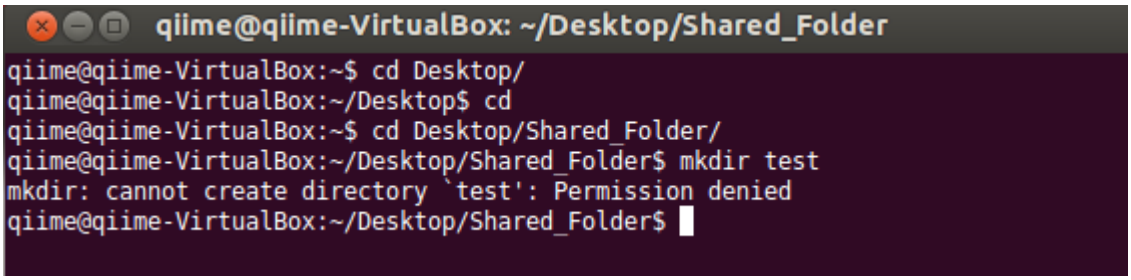
Set up the Shared Folder
The Shared Folder allows you to easily share files between the VM and the host system (your computer). Right click the gray folder in the bottom right hand corner of the VM window and select "Shared Folder" (see screenshot below).

Click the blue folder with the green plus sign to set up a new shared folder. Choose where you would like to set up the Shared Folder (Desktop or a specific folder in your file system) and rename it to **Shared_Folder**. This name is case sensitive and the underscore is important. If it is not exactly as above, VB will not recognize it as your shared folder. Check "Auto-mount" and "Make Permanent" before clicking OK. The path to the shared folder to should appear underneath Machine Folders. Click OK.



Test the Shared Folder
Click the icon in the upper right hand corner to shut down the VM. Once you are returned to the VB Manager, start the VM again. Choose the "Send Shutdown Signal" option. If the once gray folder is now blue, a shared folder is active. Once the Ubuntu desktop appears, double click on **Shared_Folder** and the contents of your shared folder should appear.

Note: You cannot write directly to the Shared Folder from the VB. You may create a new directory from the host shared folder (e.g. the desktop), but not from the Ubuntu desktop. This works in scripts as well; you cannot write or move something to the Shared Folder. If you try to do this, the error below will arise.



To get around this, create a subfolder in your Shared Folder from your host machine. You may write to this subfolder within the Shared Folder from the command line and by using the Ubuntu GUI.

The root password for the VM is **qiime**.

E. QIIME 16S WORKFLOW

1. Conventions

type text in this font directly into the command line
<use_your_own_name> (omit < >)
Comments; do not type into command line

2. Flowchart

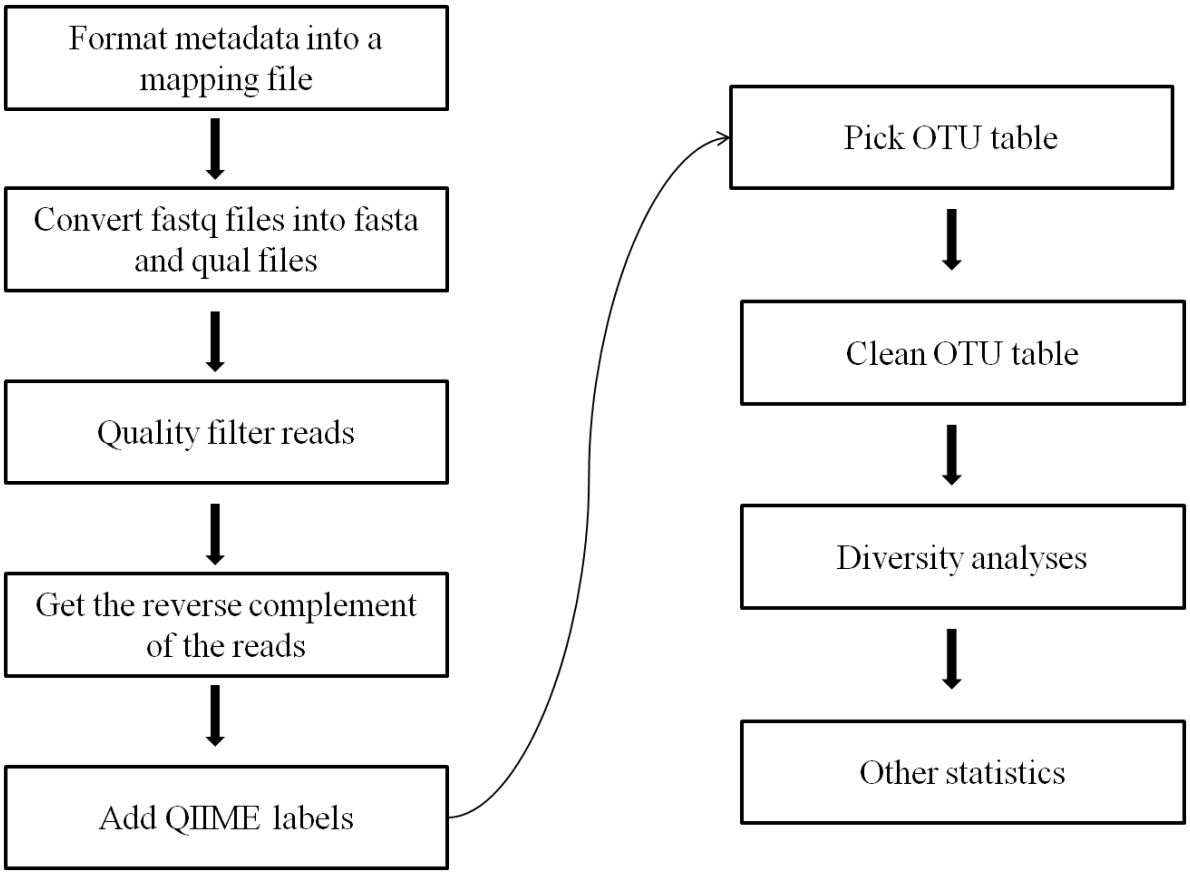

Format metadata into a mapping file
Convert fastq files into fasta and qual files
Quality filter reads
Get the reverse complement of the reads
Add QIIME labels
Pick OTU table
Clean OTU table
Diversity analyses
Other statistics

Figure 9. The QIIME workflow as presented in this tutorial.

3. Metadata

Formatting the metadata properly is challenging, but there are detailed instructions and troubleshooting tips in the QIIME documentation and on the QIIME forum. We refer to the metadata file as the **mapping file.** It is easiest to make the mapping file in Excel and save it as a tab delimited file. See the link below for more information.
http://qiime.org/documentation/file_formats.html#input-files

The essential headers are #SampleID, BarcodeSequence, LinkerPrimerSequence, and Description. The order is important, and headers are case sensitive. **All of the metadata columns are to be placed between the LinkerPrimerSequence and Description columns.**

Keep these formatting requirements in mind:

1. Sample IDs must be unique and may contain only numbers, letters, or periods.
2. Use only alphanumeric characters, underscores (except #SampleID), and periods.
3. Do not use these characters in the mapping file: $ * ^
4. Fill in blank cells with NA.
5. Avoid spaces.

Explanation of required headers

#SampleID

This column contains the unique sample names. No duplicates are allowed.

BarcodeSequence

This is the unique barcode sequence that corresponds to each unique sample ID. QIIME will match the barcoded reads to a sample ID using the information in this column. This column is required even if the sequences are already demultiplexed, for example in the case of this tutorial.

LinkerPrimerSequence

This is the linker primer pad sequence that is added to the amplicon during PCR amplification. QIIME will use this information to remove these extraneous bases from the reads. This column is required even if the sequences are already demultiplexed.

Description

This must be the last column in the mapping file. The fields must be unique. For simplicity, just copy the sample IDs from the first column into the Description column.
A sample mapping file looks like this:

The header for this mapping file starts with a pound (#) character, and generally requires a "SampleID", "BarcodeSequence", "LinkerPrimerSequence", and a "Description", all tab delimited. The following example mapping file represents the minimum field requirement for the mapping file.

#SampleID	BarcodeSequence	LinkerPrimerSequence	<metadata_1>	…	Description
sample.1	fill in the barcode	fill in the primer, etc.	values	…	sample.1
sample.2	fill in the barcode	fill in the primer, etc.	values	…	sample.2
sample.3	fill in the barcode	fill in the primer, etc.	values	…	sample.3
sample.4	fill in the barcode	fill in the primer, etc.	values	…	sample.4

Once the mapping file is created, validate the format with this handy script. Validating the mapping file prior to running any analyses will save you a lot of hassle.
http://qiime.org/scripts/check_id_map.html

A sample mapping file with common errors and omissions is included on the flash drive, as well as a correct mapping file. The correct mapping file may be used in downstream analyses if time constraints prohibit troubleshooting the mapping file with errors.

```
11-3-17-00-2-rep-2_S51_L001_R2_001.fastq.gz
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/raw_data$ cd ..
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ check_id_map.py -m
mapping.txt -o checked_mapping
No errors or warnings were found in mapping file.
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$
```

check_id_map.py
-m <mapping_file.txt>
-o <output_dir>

The output of this script is a log file, which lists the errors and warnings found in the original. A corrected mapping file will be produced, with "_" as the default replacement character. Finally, an HTML version of the mapping file will be created with the problem fields highlighted. We have found it's best to fix the errors yourself and then recheck to ensure your metadata stays as you intended.

4. Extract compressed files
The raw data is compressed into a "tarball" to save space. Extract the files using the tar utility. Move to the directory where you would like to create the raw_data directory. In this case, we are putting the raw data into a subdirectory of the Shared_Folder called cso_for_hhmi.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ tar -xzvf raw_data.
tar.gz
```

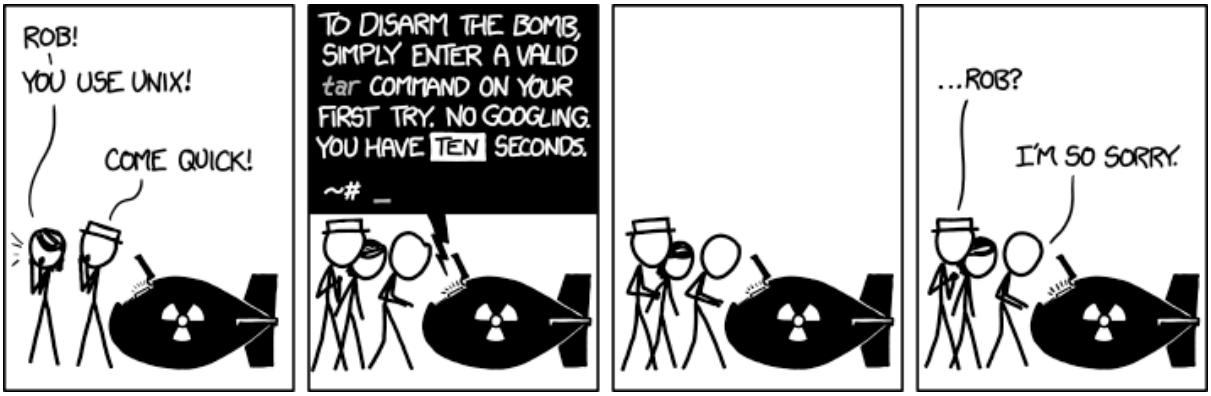
The tar command takes many options. The options we used here are explained below. See the GNU documentation of tar for more information.
http://www.math.utah.edu/docs/info/tar_2.html

-x extract files (compare –c, which is used to create an archive)
-z use the gzip utility
-v tar will print the files as they are extracted (verbose)
-f name of the archive

Move to the raw_data directory and list (ls) the contents to check. If the directory contents look like below, you are ready to proceed with the tutorial.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/raw_data$ ls
10-29-21-43-1-16-rep1_S65_L001_R2_001.fastq    10-30-9-25-3-rep1_S59_L001_R2_001.fastq
10-29-21-43-1-16-rep2_S66_L001_R2_001.fastq    10-30-9-25-3-rep2_S60_L001_R2_001.fastq
10-29-21-43-2-8-rep1_S41_L001_R2_001.fastq     10-31-0-25-1-rep2_S39_L001_R2_001.fastq
10-29-21-43-2-8-rep2_S42_L001_R2_001.fastq     10-31-0-25-1_S50_L001_R2_001.fastq
10-29-21-43-3-7_S36_L001_R2_001.fastq          10-31-0-25-2-rep1_S53_L001_R2_001.fastq
10-30-12-37-1-10_S37_L001_R2_001.fastq         10-31-0-25-2-rep2_S54_L001_R2_001.fastq
```

Now that you have successfully extracted the archive, let's break for some relevant humor.


ROB!
YOU USE UNIX!
COME QUICK!
TO DISARM THE BOMB, SIMPLY ENTER A VALID tar COMMAND ON YOUR FIRST TRY. NO GOOGLING. YOU HAVE TEN SECONDS.
~# _
...ROB?
I'M SO SORRY.

http://xkcd.com/1168/

You will likely generate a number of large files in the process of analyzing metagenomic data. Therefore, you will likely come across tar more often in the future. Unfortunately, it's nearly impossible to keep the flags straight, especially across Unix/Linux distributions and zipping utilities!

5. Split Libraries Workaround

Since we are dealing with pre-demultiplexed data in fastq format, it is necessary to convert the fastq files to fasta and qual files before running any other scripts, as these scripts can only handle the fasta file format. We can use bash scripts to automate these steps to save typing and reduce the possibility of errors. Bash (Bourne again shell) scripts are small programs written in the terminal, which you can use to automate certain tasks. While bash scripting takes some practice, implementing while loops can save a lot of time.

a. Make a list of the files you want to convert from fastq to fasta/qual format.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ cd raw_data/
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/raw_data$ ls > list.t
xt
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/raw_data$ gedit list.
txt
```

cd raw_data

Move to the directory where your unzipped raw data is located.

ls > list.txt

Write the output of the list to a text file.

gedit list.txt

Open the text file with gedit, a Notepad-like text editor. Ensure that only the files you want to unzip are in this file. You will probably have to delete list.txt and a blank line after the files. Once this is done, save and close gedit. You will return to the command prompt.

b. Convert fastq files to fasta files using a bash script.
While there is a line to read, the file specified in the line will be converted from a fastq file to a fasta file and a qual file and written to converted_files. After all files have been converted, the list will close and the command prompt will return. This step may take up to ten minutes.
http://qiime.org/scripts/convert_fastaqual_fastq.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/raw_data$ while read line; do time convert_fa
staqual_fastq.py -f $line -o /home/qiime/Desktop/Shared_Folder/cso_for_hhmi/converted_files/ -c fastq_to_
fastaqual;  done < list.txt
```

You can either type everything at once or hit enter after the semicolons.

while read line;
do time convert_fastaqual_fastq.py –f $line –o converted_files/ -c fastq_to_fastaqual;
done < list.txt

Remove the list.
rm list.txt*

c. Generate quality plots

This script requires a lot of memory, which is probably limited on personal computers. Use GCAT-SEEK, which has enough memory for this script to complete in an hour. Alternatively, see the complete output in qual_plots.tar.gz to save time. The output of this script is not required to proceed with the tutorial.

Reads from the sequencing instrument can be of differing quality, and it's important to ensure that the reads you are working with are of good quality to improve the downstream results. Phred scores indicate the probability that a base call is correct. For example, if a base has a Phred score of 20, the chance that this base call is correct is 99%. Phred scores are calculated with a logarithm, so a Phred score of 30 indicates that the probability of a correct base call is 99.9% for a certain position.

QIIME allows you to determine at which nucleotide the quality tends to drop off. Using this information, you can truncate the reads manually or appropriately set a quality cutoff in split_libraries_fastq.py.
http://qiime.org/scripts/quality_scores_plot.html

Again, we can automate this process with a bash script. Move to the directory containing the fasta and qual files. Make a new directory to store the qual files and copy them there. Then make a list of the qual files as before and remove the files you do not want.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files$ cd
qiime@qiime-VirtualBox:~$ cd Desktop/Shared_Folder/cso_for_hhmi/converted_files/
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files$ mkdir qual_files
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files$ cp *.qual qual_files
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files$ cd qual_files/
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/qual_files$ ls > qual_list.txt
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/qual_files$ gedit qual_list.txt
```

cp *.qual qual_files

cd qual_files
ls > qual_list.txt
gedit qual_list.txt

Then, implement the loop. This script generates plots of sequence quality and returns the position where the quality falls below the Phred score specified with –s. The default Phred quality score is 25; however we use 20 in this example.

Make an output directory with the utility mkdir.
mkdir /home/qiime/Desktop/Shared_Folder/cso_for_hhmi/qual_plots

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/qual
_files$ cd
qiime@qiime-VirtualBox:~$ cd Desktop/Shared_Folder/cso_for_hhmi/converted_files/
qual_files/
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/qual
_files$ while read line;
> do time quality_scores_plot.py -q $line -s 20 -o /home/qiime/Desktop/Shared_Fo
lder/cso_for_hhmi/qual_plots/$line;
> done < qual_list.txt
```

Move to the directory containing the qual files and implement the loop. For each qual file, position at which the average Phred score falls below 20 is recorded and a graph of the average Phred scores over the length of the reads is created.

cd /home/qiime/Desktop/Shared_Folder/cso_for_hhmi/converted_files/qual_files/

while read line;
do time quality_scores_plot.py -q $line –s 20 -o
/home/qiime/Desktop/Shared_Folder/cso_for_hhmi/qual_plots/$line/;
done < qual_list.txt

Remove the list of quality files and the qual_files directory.
rm qual_list.txt*
rm -R qual_files

The output will be contained in separate directories for each sample. The PDF file contains a plot of the average quality score at each nucleotide for all of the reads in each sample. The dotted line indicates the minimum quality score, which we specified as Phred 20. The black line is the average quality score, with the upper and lower red lines indicate the standard deviation. In this case we have good quality along the length of the read, since the average quality score falls below Phred 20 just before the 250th nucleotide position.

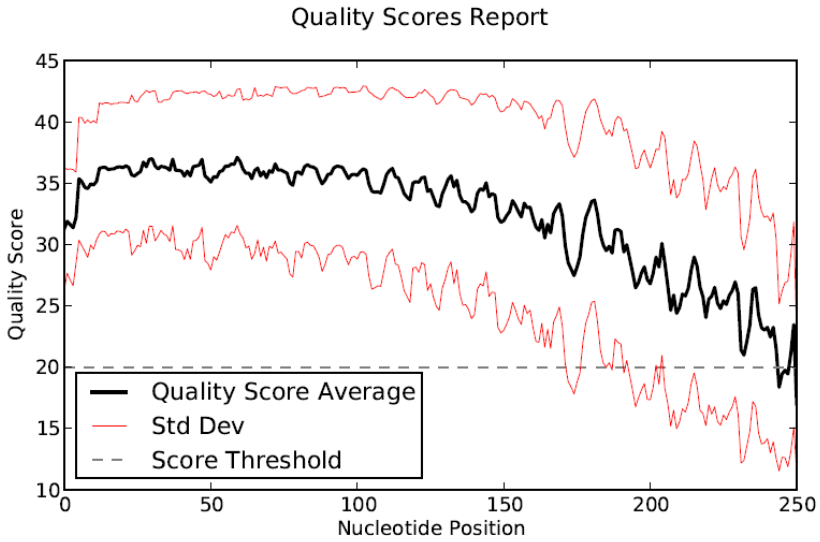


Figure 10. A quality scores report shows the average quality score per nucleotide position. This analysis is used for determining where to truncate the reads to remove regions of poor quality.

We can verify this by looking at the text file output. The header to the text file, which is written to the same directory as the quality plot, suggests that the reads in this sample should be truncated to 244 nt. We have 250 bp reads, so losing only 6 bp is very good.

Suggested nucleotide truncation position (None if quality score average did not drop below the score minimum threshold): **244**
… more raw data …

d. Truncate the files. After reviewing all of the data, we can pick a position at which to truncate the reads. We want to maximize the quality of the reads, so let's truncate at position 220 to preserve as many quality sequences as possible. We can use a bash script to automate this process, as we have done above. **This script takes approximately 25 minutes to complete.**
http://qiime.org/scripts/truncate_fasta_qual_files.html

Move to the directory where the converted fasta and qual files are contained. As before, write a list of the contents to a text file. Use gedit to remove any unwanted files. Save and close to return to the command prompt.

cd converted_files/
ls > conv_files.txt
gedit conv_files.txt

Implement the loop, which will read in the fna and qual files from the list and truncate each file at nucleotide 220. The output will be written to the directory specified with –o.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files$ whi
le read fna;
> do read qual;
> time truncate_fasta_qual_files.py -f $fna -q $qual -b 220 -o truncated_fasta/;
> done < conv_files.txt
```

while read fna;
do read qual;
time truncate_fasta_qual_files.py -f $fna -q $qual -b 220 -o truncated_fasta/;
done < conv_files.txt

Remove the unneeded list.
rm conv_files.txt*

e. Generate the reverse complement of the sequences

The reads used in this tutorial are from a paired end sequencing run. Paired end sequencing allows reads both in the forward and reverse directions to be generated. This is useful when attempting to sequence a long amplicon with short read lengths. Support for paired end reads is currently limited in QIIME, so typically single end reads are used to generate the OTU table. We are using read 2 because the quality is much better than read 1 in this case. The read 2's are in reverse orientation relative to the Greengenes database; therefore we must generate the reverse complement of the reads before picking the OTU table. As before, we can use a bash script to automate the process.
http://qiime.org/scripts/adjust_seq_orientation.html

Move to the directory containing the truncated fasta files. Make a new directory to store the fasta files temporarily. Then, copy only the fasta files to this new directory. List the contents and write to a file. Use gedit to remove files that should not be reversed.

```
qiime@qiime-VirtualBox:~$ cd Desktop/Shared_Folder/cso_for_hhmi/converted_files/
truncated_fasta/
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/trun
cated_fasta$ mkdir trunc_fasta_only
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/trun
cated_fasta$ cp *.fasta trunc_fasta_only/
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/trun
cated_fasta$ cd trunc_fasta_only/
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/trun
cated_fasta/trunc_fasta_only$ ls > trunc_fasta.txt
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/trun
cated_fasta/trunc_fasta_only$ gedit trunc_fasta.txt
```

cd <truncated_fasta>
mkdir <trunc_fasta_only>
cp *.fasta trunc_fasta_only/
cd trunc_fasta_only
ls > trunc_fasta.txt
gedit trunc_fasta.txt

Make an output directory for the reverse complements of the truncated sequence files. Implement the loop. Make sure the working directory contains only the truncated fasta files.

```
qiime@qiime-VirtualBox:~$ mkdir Desktop/Shared_Folder/cso_for_hhmi/rc_seqs/
qiime@qiime-VirtualBox:~$ cd Desktop/Shared_Folder/cso_for_hhmi/converted_files/
truncated_fasta/trunc_fasta_only/
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/converted_files/trun
cated_fasta/trunc_fasta_only$ while read line;
> do time adjust_seq_orientation.py -i $line -o /home/qiime/Desktop/Shared_Folde
r/cso_for_hhmi/rc_seqs/$line;
> done < trunc_fasta.txt
```

mkdir rc_seqs/
while read line;
do time adjust_seq_orientation.py –i $line –o
/home/qiime/Desktop/Shared_Folder/cso_for_hhmi/rc_seqs/$line
done < trunc_fasta.txt

Remove the unneeded list.
rm trunc_fasta.txt*

f. Add QIIME labels. This script allows you to specify the QIIME compatible label that belongs to each file. Since we can't use split libraries as described in the documentation, we have to add the labels manually. This script requires the fasta files and a modified mapping file with the desired QIIME label and the exact name of the corresponding file.

http://qiime.org/scripts/add_qiime_labels.html

Modify the mapping file such that the file name is correct. The FILE_NAME must match the file to be labeled. The SampleID can be anything within the formatting guidelines, but it is often helpful to keep it as simple as possible.

#SampleID	other	FILE_NAME
S1	metadata	10-29-21-43-1-16-rep1_S65_L001_R2_001_filtered.fasta
S2	…	10-29-21-43-1-16-rep2_S66_L001_R2_001_filtered.fasta
S3	…	10-29-21-43-2-8-rep1_S41_L001_R2_001_filtered.fasta
S4	…	10-29-21-43-2-8-rep2_S42_L001_R2_001_filtered.fasta

Add the QIIME labels. The script requires a directory as input, so we don't have to use a bash script here to add the labels to every sequence.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ add_qiime_labels.py
 -m mapping_files/mod_mapping.txt -i rc_seqs/ -c FILE_NAME -n 1000000 -o labeled
_rc_seqs/
```

add_qiime_labels.py
-m < mapping.txt>
Don't forget to put in the file path.
-i <fasta_dir>
Ensure that only the desired fasta files are in this directory.
-c <name>
This is the header of the column in the mapping file where the file names are located.
-n 1000000

Each read needs a unique identifier. Starting at 1000000 is recommended.
-o <output_dir>

The output of this script will be one fasta file called combined_seqs.fna containing all of the reads with QIIME labels.

g. Validate the QIIME labels. Before moving on, it's good to check that the labels were added correctly and that QIIME can understand the sample names. If this script fails, you will not be able to do any of the downstream analyses.
http://qiime.org/scripts/validate_demultiplexed_fasta.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi/labeled_rc_seqs$ val
idate_demultiplexed_fasta.py -i combined_seqs.fna -o /home/qiime/Desktop/Shared_
Folder/cso_for_hhmi/validated_fasta/ -m /home/qiime/Desktop/Shared_Folder/cso_fo
r_hhmi/mapping_files/mod_mapping.txt
```

validate_demultiplexed_fasta.py
-i combined_seqs.fna
-m <mapping_file.txt>
-o <output_dir>

A log file will be written to the output directory. Make sure there are no QIIME – incompatible fasta labels.

fasta file combined_seqs.fna validation report
Percent duplicate labels: 0.000
Percent QIIME-incompatible fasta labels: 0.000
Percent of labels that fail to map to SampleIDs: 0.000
Percent of sequences with invalid characters: 0.000
Percent of sequences with barcodes detected: 0.000
Percent of sequences with barcodes detected at the beginning of the sequence: 0.000
Percent of sequences with primers detected: 0.000

6. OTU Table Picking
Picking the OTU table is the most computationally intensive step in the entire workflow. Fortunately, there exist a few ways to reduce the computational time needed while retaining most of the informative reads. Closed reference, *de novo*, and open reference OTU picking are three ways OTU picking can be done in QIIME.

We use the Greengenes 16S rRNA database, which is pre-clustered and of high quality (chimera checked). Approximately annual updates are released. It is advisable to use the latest version of the database, especially if you are characterizing non-human associated environments, since thousands of new OTUs are added with each update.

Closed reference OTU picking uses the algorithm UCLUST to compare the reads against Greengenes. If a read fails to match a reference sequence at least 70% identity, it is discarded. Closed reference OTU picking is comparably fast; however many reads are discarded. We will use closed reference OTU picking in this workshop.

***De novo* OTU** picking is much slower; however no reads are discarded. The reads are clustered with themselves and assigned taxonomies. Chimeric sequences could be present, so it is essential to check for chimeras before proceeding with downstream analyses.

Open reference OTU picking is a compromise between closed reference and *de novo* OTU picking. Initially, the reads are compared to the reference database, and those that fail to match a reference sequence at at least 70% identity are retained. The reads that didn't hit the database are clustered *de novo* and taxonomies are assigned. Open reference OTU picking is slower than closed reference; reads that may represent novel OTUs are not discarded. The QIIME developers recommend open reference OTU picking.

http://qiime.org/scripts/pick_closed_reference_otus.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/erin_cso$ pick_closed_reference_o
tus.py -i combined_seqs.fna -a -O 15 -o uclust_closed_otus/ -r /home/qiime/qiime
_software/gg_13_5_otus/rep_set/97_otus.fasta -t /home/qiime/qiime_software/gg_13
_5_otus/taxonomy/97_otu_taxonomy.txt
```

pick_closed_reference_otus.py
-i <input_seqs.fna>
The sequences **must** be in fasta format
-o <output_dir>
-r <rep_set>
Use the representative set from the latest version of Greengenes. We typically use 97%, as OTUs with 97% identity very roughly correspond to the same species.
-t <taxonomy_map>
Use the taxonomy from the latest version of Greengenes
-a
run in parallel
-O <num_jobs_to_start>

7. Initial Analyses

a. OTU Table Statistics
This script summarizes the OTU table information, which allows you to find out basic information about the OTU table to verify that OTU picking occurred as expected.
http://biom-format.org/documentation/summarizing_biom_tables.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ print_biom_table_su
mmary.py -i otu_table.biom
Num samples: 46
Num observations: 19495
Total count: 1529520.0
Table density (fraction of non-zero values): 0.1772
Table md5 (unzipped): 1b69e8f8b01cf532e4e8335bdb67d0d5

Counts/sample summary:
 Min: 2.0
 Max: 184870.0
 Median: 12409.5
 Mean: 33250.4347826
 Std. dev.: 46267.857757
 Sample Metadata Categories: None provided
 Observation Metadata Categories: taxonomy

Counts/sample detail:
 S36: 2.0
 S27: 3.0
 S6: 3.0
 S7: 17.0
 S47: 47.0
```

print _biom_table_summary.py
-i otu_table.biom
--num_observations
Optional; returns the number unique OTUs in each sample.

The results of this script are written to the terminal. We can see that there are 46 samples in the OTU table. We have 19,495 OTUs (Num observations), and a total of 1,529,250 reads. In Count/sample detail the number of sequences in each sample is listed in ascending order.

b. Clean the OTU Table
If it suits your study, it is a good idea to remove OTUs with less than 10 observances. The OTU table should also be rarefied prior to doing any further analysis, as even sequencing depth is required to compare samples. The final OTU table should be used in all downstream analyses, except in alpha rarefaction and when using the beta diversity workflow.
http://qiime.org/scripts/filter_otus_from_otu_table.html
http://qiime.org/scripts/single_rarefaction.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ filter_otus_from_ot
u_table.py -i otu_table.biom -o otu_table_min10.biom -n 10
```

filter_otus_from_otu_table.py
-i <rarefied_otu_table.biom>
-o <output_otu_table.biom>
-n 10

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ single_rarefaction.
py -i otu_table_min10.biom -d 1000 -o otu_table_min10_even1000.biom
```

single_rarefaction.py
-i otu_table.biom

-o <output_otu_table.biom>
-d 1000

When choosing a depth at which to rarefy, it is important to maximize the number of samples kept while avoiding samples with a very low number of sequences. A good rule of thumb is to exclude samples with fewer than 1000 seqs/sample.

c. Summarize Taxa

The OTU table is divided into taxonomic level and either the relative abundance or absolute abundance of each taxon in each sample is returned. These taxa summaries are useful to detect broad trends in the dataset and to make bioplots (see beta diversity; make _3d_plots.py). There is a workflow script for generating plots; however we have found it easier to manipulate the data in Excel.
http://qiime.org/scripts/summarize_taxa.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ summarize_taxa.py -
i otu_table_even1000_min10otus.biom -o summarized_taxa/
```

summarize_taxa.py
-i <clean_otu_table.biom>
-o <output_dir>
The default ouput of this script are taxa summaries in both tab delimited and biom formats. Taxa are summarized by L2 = phylum, L3 = class, L4 = order, L5 = family and L6 = genus. Taxa can also be summarized at the species level (L7); however the V4 region of the 16S rRNA gene typically does not provide enough resolution to get species level taxonomic information. See the documentation for more information.

An excerpt from a family level (L5) taxa summary is shown here. The taxa are in column A, with the relative abundance of each taxon given per sample, which are labeled in row 1.

	A	B	C	D	E	F	G
1	Taxon	S2	S19	S5	S18	S15	S12
2	k__Archaea; p__Crenarchaeota; c__Thaumarchaeota; o__Cenarchaeales; f__Cenarchaeaceae	0	0	0	0.003175	0.001642	
3	k__Archaea; p__Crenarchaeota; c__Thaumarchaeota; o__Cenarchaeales; f__SAGMA-X	0	0	0.001534	0	0.001642	
4	k__Bacteria; p__Acidobacteria; c__Acidobacteriia; o__Acidobacteriales; f__Acidobacteriaceae	0.001499	0.004451	0.001534	0	0	0.001
5	k__Bacteria; p__Acidobacteria; c__Sva0725; o__Sva0725; f__	0	0	0	0.001587	0.001642	0.003
6	k__Bacteria; p__Acidobacteria; c__iii1-8; o__DS-18; f__	0.002999	0.005935	0.001534	0.012698	0.00821	0.006
7	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Kineosporiaceae	0	0	0	0.001587	0.001642	
8	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Microbacteriaceae	0.002999	0.008902	0.003067	0.014286	0.006568	0.001
9	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Micrococcaceae	0	0	0	0	0	

Figure 11. The text files output by the summarize taxa script shows the relative abundance of each taxon (at different taxonomic levels) in each sample.

8. Diversity Analyses

Microbial ecologists are often interested in the microbial diversity of their samples. Alpha diversity refers to the within sample diversity, while beta diversity describes the differences in diversity between samples. Both alpha and beta diversity can be computed in QIIME, and a number of metrics for each measure of diversity can be implemented.

a. Alpha Diversity

While QIIME offers a workflow script for generating rarefaction plots, previous experience has shown that it often fails and does not progress past the multiple rarefactions. Therefore, it is advisable to do each step of the workflow separately.

Alpha diversity is computed by generating multiple rarefactions of the OTU table at different sampling depths, calculating alpha diversity on each rarefied OTU table, collating the results, and making rarefaction plots. The most computationally intensive step of this workflow is performing the multiple rarefactions.

Refer to the QIIME documentation for more information and other options.

i. Generate multiple rarefactions
Run this step in parallel if possible. For serial environments, use multiple_rarefactions.py and omit –O. Use the OTU table with OTUs observed at least 10 times in the dataset. This script will take approximately 20 minutes to complete using the options below (3 cores, 3 GB RAM).
http://qiime.org/scripts/parallel_multiple_rarefactions.html

This script will yield 100 OTU tables in total. The original OTU table will be subsampled 10 times at a depth of 10 seqs/sample, 10 times at a depth of 120 seqs/sample, and so on until the maximum rarefaction depth is reached (1000 seqs/sample). The step size is 110, which means that the each sampling depth will be increased by 110 until 1000 seqs/sample is reached. Since each subsampling undergoes 10 iterations, a total of 100 subsampled OTU tables will be generated.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ parallel_multiple_r
arefactions.py -i otu_table_min10.biom -o multiple_rarefactions/ -m 10 -x 1000 -
n 10 -s 110 -O 3
```

parallel_multiple_rarefactions.py
-i otu_table.biom
-o <output_directory>
Put the absolute path of the output directory where you want to store the rarefied OTU tables.
-m <min_rarefaction_depth>
-x <max_rarefaction_depth>
-s <step_size>
-O <num_jobs_to_start>
Maximize the number of jobs to start, since this step is computationally intensive.

The output of this script is a directory containing 10 OTU tables per sampling depth.

ii. Compute alpha diversity
Run this step in parallel if possible. For serial environments, use alpha_diversity.py and omit –O. This step takes about 20 minutes using the metrics below (3 cores, 3 GB RAM).
http://qiime.org/scripts/parallel_alpha_diversity.html
http://qiime.org/scripts/alpha_diversity_metrics.html

We will use the alpha diversity metrics chao1, PD_whole_tree, and Heip's evenness to calculate the richness and evenness of the dataset.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ parallel_alpha_dive
rsity.py -i multiple_rarefactions/ -o alpha_div/ -t /home/qiime/qiime_software/g
g_13_5/gg_13_5_otus/trees/97_otus.tree -m chao1,PD_whole_tree,heip_e -O 3
```

parallel_alpha_diversity.py
-i <input_dir>
Input the directory containing the multiple rarefactions.
-o <output_dir>
-t <path to reference tree file>
Use the latest Greengenes tree. Needed for PD_whole_tree only
-m <alpha_div_metrics>
Separate multiple metrics with commas; do not include spaces.
-O <num_jobs_to_start>

This script will output tab delimited files with the results of each alpha diversity metric at each sampling depth. Sometimes the script completes without deleting the temporary folders. If this is the case, the next script will fail. If the alpha diversity results look complete and all iterations are a similar size, try to delete the temporary directories before collating.

iii. Collate alpha diversity results
http://qiime.org/scripts/collate_alpha.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ collate_alpha.py -i
 alpha_div/ -o collated_alpha/
```

collate_alpha.py
-i <input_dir>
Input the directory containing the results of parallel_alpha_diversity.py
-o <output_dir>

The output of this script is one text file per metric containing the alpha diversity results.

iv. Plot rarefaction curves
See the other optional parameters for adjusting image resolution or file type. This script may take a few minutes to complete.
http://qiime.org/scripts/make_rarefaction_plots.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ make_rarefaction_pl
ots.py -i collated_alpha/ -m mapping_files/mapping_complete.txt -o rarefaction_p
lots/
```

make_rarefaction_plots.py
-i <input_dir>
Input the directory containing the results of collate_alpha.py
-m <mapping_file>
-o <output_dir>

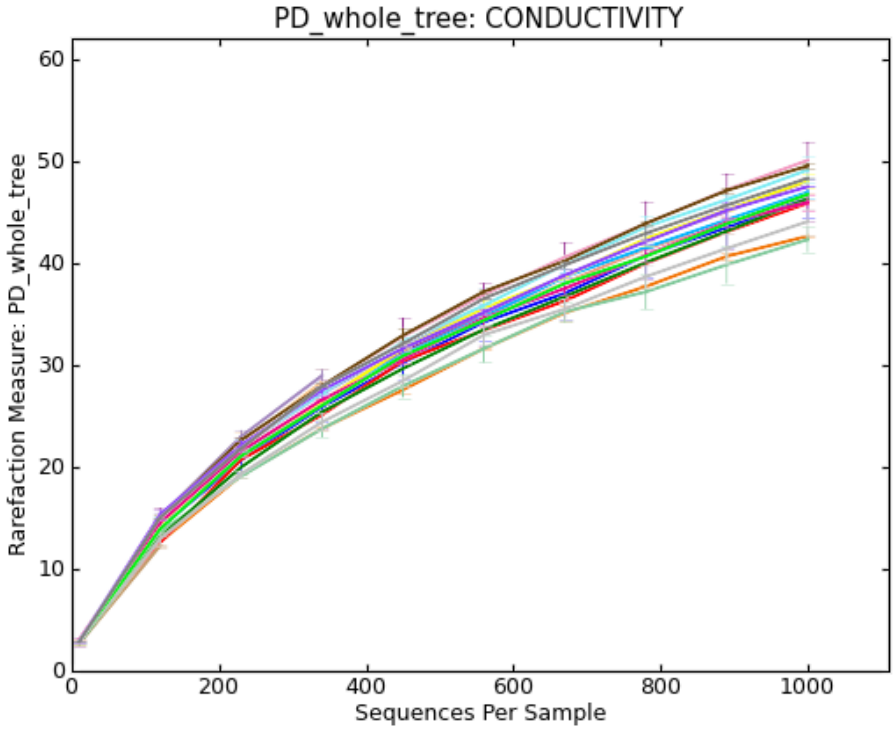


Figure 12. A sample alpha rarefaction plot showing the alpha diversity of the OTU table over a range of sampling depths.

In this rarefaction plot, the sampling depth is plotted on the x axis, while the species richness calculated with the alpha diversity metric PD whole tree is on the y axis. Each color represents a different conductivity value. Rarefaction plots can reveal differences in alpha diversity between treatments or timepoints; however, here it seems that the alpha diversity at each conductivity value is similar even as sampling depth increases. Rarefaction plots can also tell you how well sampled the environment is. If the OTU richness increases as more OTUs are being sampled, then you probably should rarefy to a deeper sampling depth. In this case, species richness drops off as sampling depth increases. The HTML output file includes all of the metadata categories and metrics used, so it is easy to look at alpha diversity as calculated by multiple metrics and colored by multiple metadata variables. The script compare_alpha.py will allow you to determine if differences in alpha diversity values are significant.

b. Beta Diversity

QIIME offers a workflow script for computing beta diversity, or each step can be performed individually. The current workflow script uses KiNG to visualize the 3D PCoA plots; however splitting up the workflow into individual scripts allows you to use Emperor, a program specifically written by the Knight Lab to visualize PCoA plots.

The most computationally intensive part of beta diversity calculations is computing the distance matrix of pairwise dissimilarity values between samples, so be sure to run this script in parallel if possible. There is a number of different beta diversity metrics supported in QIIME; however we have found the Unifrac distance metrics to be most useful and informative. Unifrac takes phylogenetic relatedness into account in computing beta diversity. Unweighted Unifrac regards only the presence or absence of taxa, whereas weighted Unifrac uses taxon relative abundance as well.

Workflow script
http://qiime.org/scripts/beta_diversity_through_plots.html

beta_diversity_through_plots.py
-i otu_table.biom
-m <mapping_file>
-o <output_dir>
-t < path_to_ref_tree>
Use the latest Greengenes reference tree; required for Unifrac
-a
required to run in parallel
-O <num_jobs_to_start>
-e <seqs_per_sample>
depth of coverage for even sampling

Individual Scripts
i. Compute distance matrices
This script takes about 7 minutes to run on three cores with about 3 GB of RAM.
http://qiime.org/scripts/parallel_beta_diversity.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ time parallel_beta_
diversity.py -i otu_table_min10_even1000.biom -t /home/qiime/qiime_software/gg_1
3_5/gg_13_5_otus/trees/97_otus.tree -O 3 -o beta_div_dms
```

parallel_beta_diversity.py
-i <clean_otu_table.biom>
-o <output_dir>
-m <metrics>
Default metrics are unweighted and weighted Unifrac. Additional metrics are given at http://qiime.org/scripts/beta_diversity_metrics.html
-t <path_to_ref_tree>
Use the latest Greengenes reference tree
-O <num_jobs_to_start>

The output of this script is two distance matrices, one of which is calculated with the unweighted Unifrac metric, and the other which is calculated with weighted Unifrac. An excerpt of the unweighted distance matrix is included below. Samples are compared pairwise to all other samples. The values are the same across the diagonal due to the pairwise comparisons, e.g. B3 and C2 are identical since both S9 and S8 are being compared.

	A	B	C	D	E	F	G	H	I	J	K
		S9	S8	S3	S2	S1	S5	S4	S19	S18	S39
2	S9	0	0.417418	0.418494	0.505656	0.436287	0.472299	0.476023	0.393154	0.454499	0.405428
3	S8	0.4174	0	0.433982	0.443292	0.470943	0.461472	0.468011	0.439557	0.443821	0.510988
4	S3	0.4185	0.433982	0	0.435458	0.43856	0.463217	0.407056	0.446045	0.51256	0.50475
5	S2	0.5057	0.443292	0.435458	0	0.444524	0.441994	0.482498	0.497162	0.506902	0.565564
6	S1	0.4363	0.470943	0.43856	0.444524	0	0.396904	0.482262	0.496	0.465256	0.532358
7	S5	0.4723	0.461472	0.463217	0.441994	0.396904	0	0.411945	0.443405	0.466943	0.531708
8	S4	0.476	0.468011	0.407056	0.482498	0.482262	0.411945	0	0.442192	0.531609	0.525095

Figure 13. Sample unweighted Unifrac distance matrix.

ii. Compute principal coordinates matrices. Do this for both unweighted and weighted distance matrices.
http://qiime.org/scripts/principal_coordinates.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ mkdir beta_div_pc
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ principal_coordinat
es.py -i beta_div_dms/unweighted_unifrac_otu_table_min10_even1000.txt -o beta_di
v_pc/unweighted_pc.txt
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ principal_coordinat
es.py -i beta_div_dms/weighted_unifrac_otu_table_min10_even1000.txt -o beta_div_
pc/weighted_pc.txt
```

principal_coordinates.py
-i <distance_matrix.txt>
-o <pc_matrix.txt>

The output of this script is a text file with the distance matrix transformed into principal coordinates, or the combinations of variables that most explain the variation in the dataset. The principal coordinates matrix is used to generate 2D and 3D plots, which allow you to determine if any clustering of the samples is explained by a metadata category.

	A	B	C	D	E	F	G	H	I	J
1	pc vector	1	2	3	4	5	6	7	8	9
2	S9	-0.01642	0.049978	-0.02439	0.118965	-0.03807	0.027451	-0.11092	-0.0258	0.045208
3	S8	0.124188	-0.05052	0.012989	-0.03037	-0.08119	-0.03197	-0.06518	-0.1011	-0.04757
4	S3	0.121453	0.204422	-0.06203	-0.01384	-0.03246	-0.06647	-0.04391	-0.10829	-0.00974
5	S2	0.227945	0.059323	-0.10107	-0.04393	-0.19171	0.088021	0.082157	0.020069	-0.02187
6	S1	0.139404	0.025643	-0.20816	0.004351	0.049556	0.082484	-0.09065	-0.0077	0.036441
7	S5	0.172088	0.00131	-0.09028	-0.06313	0.130354	0.080865	0.083172	0.073391	0.074349
8	S4	0.143175	0.184739	-0.01257	-0.05678	0.10142	-0.10056	0.06921	0.057695	-0.12751

Figure 14. Sample principal coordinates analysis of the unweighted Unifrac distances.

iii. Make the preferences files
This script allows you to specify the coloring of the 2D and 3D plots. In this case, we will use it to get continuous coloring for the 2D plots to make it easier to observe gradients in metadata variables like salinity.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ make_prefs_file.py
-m mapping_files/mapping_complete.txt -o prefs_file.txt
```

make_prefs_file.py
-i <mapping_file.txt>
-o <prefs_file.txt>

The output of this script is a text file specifying the coloring for each metadata variable.

iv. Generate 2D plots. If you have a large mapping file, this script might fail because of memory issues. If so, you can choose to color by specific categories or split up the mapping file. See the documentation for more information.
http://qiime.org/scripts/make_2d_plots.html

Make 2D plots for both the unweighted and weighted principal coordinates matrices.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ time make_2d_plots.
py -i beta_div_pc/unweighted_pc.txt -m mapping_files/mapping_complete.txt -o 2d_
out/ -p prefs_file.txt
```

make_2d_plots.py
-i <pc_matrix.txt>
-m <mapping_file>
-o <output_dir>
-p <prefs_file.txt>

The output of this script is an HTML file containing the 2D plots. Hover over each point to see the sample ID and the metadata value. An example of a 2D plot based on unweighted Unifrac distances is shown below.

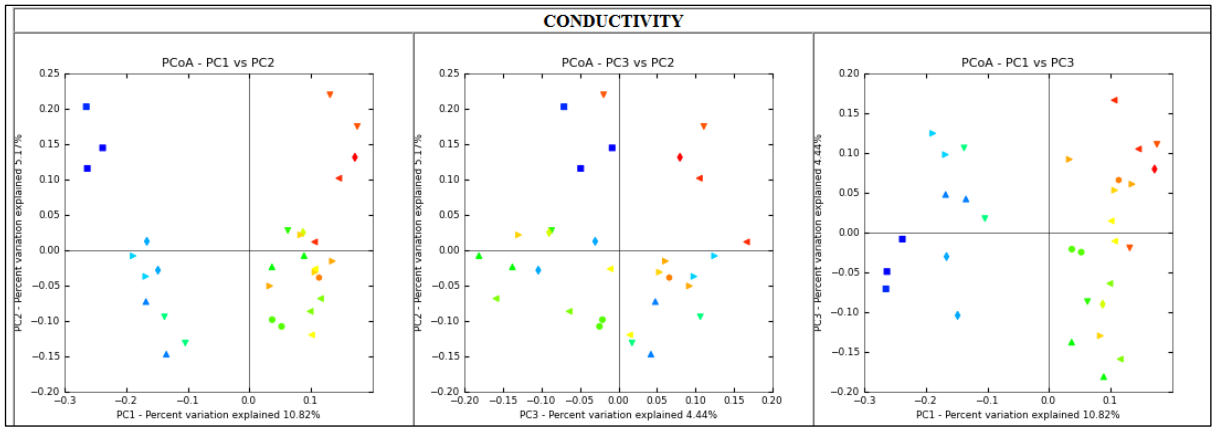

CONDUCTIVITY
PCoA - PC1 vs PC2
PCoA - PC3 vs PC2
PCoA - PC1 vs PC3
PC1 - Percent variation explained 10.82%
PC2 - Percent variation explained 5.17%
PC3 - Percent variation explained 4.44%

Figure 15. A sample 2D principal coordinates analysis colored continuously by conductivity. Hovering over a sample in the interactive HTML plot shows the sample ID and the metadata value. In this case, samples with low conductivity are located on the right (red), while high conductivity samples are located on the left (blue) in PC1 vs. PC2.

For the metadata variable conductivity, the points are colored by conductivity values on a color scale from red (low) to blue (high). The first PCoA plot is of PC1 and PC2, or the two axes which explain the most variation in the dataset. PC1 explains 10.82% and PC2 explains 5.17%. The next PCoA plots show PC2 vs. PC3 and PC1 vs. PC3, respectively.

We typically look at the first PCoA plot (PC1 vs PC2) because the most variation in the dataset is explained by this ordination of samples based on their unweighted Unifrac distances. Samples that are closer together represent communities with more similar microbial communities than do two samples that are farther apart. Furthermore, we can see here that there is a gradient of conductivity values in the sample ordination. Samples with low conductivity cluster on the right, with increasing conductivity values toward the left side of

the PCoA plot. The sample ordination in space is random, so the direction of the gradient may flip if this plot is remade; however, the trends and clustering are preserved.

v. Generate 3D plots

QIIME 1.7.0 offers support for two types of 3D plot viewers. Emperor was developed by the Knight Lab and is more user-friendly than KiNG, which was originally written for structural biology research. There are two different scripts for making 3D plots. Generate both of them and see which is more useful for your study. Make sure to make 3D plots for both unweighted and weighted Unifrac principal coordinates matrices.
http://qiime.org/scripts/make_emperor.html
http://qiime.org/scripts/make_3d_plots.html

make_emporer.py (for Emperor)

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ make_emperor.py -i
beta_div_pc/unweighted_pc.txt -m mapping_files/mapping_complete.txt -o 3d_unweig
hted_emperor/
```

OR

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ make_3d_plots.py -i
 beta_div_pc/unweighted_pc.txt -m mapping_files/mapping_complete.txt -o 3d_unwei
ghted_king
```

make_3d_plots.py (for KiNG)
-i <pc_matrix.txt>
-m <mapping_file>
-o <output_dir>

The 3D plots can be customized by adding vectors to connect samples (--add_vectors), defining an explicit axis (ex: -a pH), and/or by making biplots (-t to input taxa summaries). See the documentation for more information.

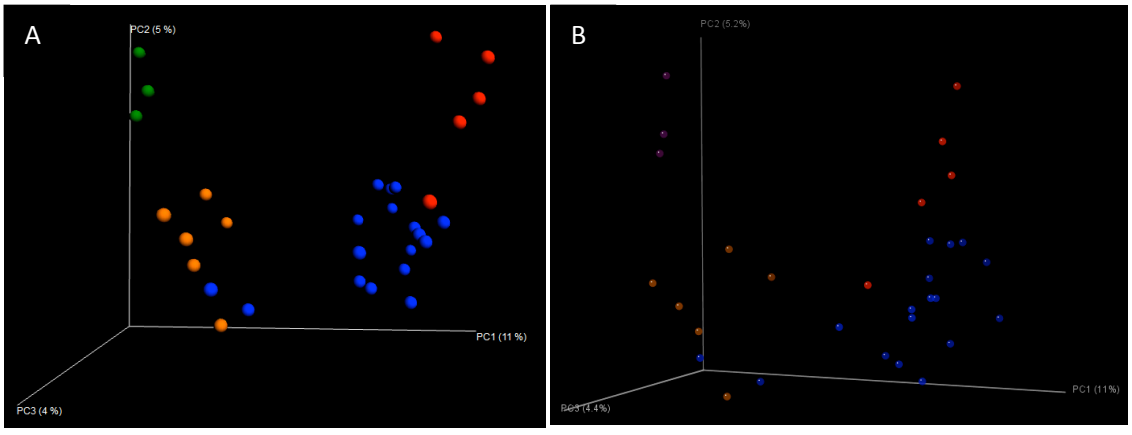


Figure 16. Sample 3D principal coordinates analysis of unweighted Unifrac distances. Plot A was generated with Emperor, while Plot B was generated with KiNG. Samples are colored by date.

The outputs of these scripts are HTML files with the 3D plots. Both Emperor (Chrome and Firefox) and KiNG run in web browers; however KiNG requires Java, which doesn't always cooperate. Below is an example of a 3D plot in Emperor (A) and in KiNG (B).
The samples are colored by date in both plots. Coloring by sample date seems to delineate the sample clustering by unweighted Unifrac distances fairly well. Samples in red were taken on 10/29, in blue on 10/30, in orange on 10/31, and in green (purple) on 11/03. While both plots yield the same information, the rendering quality is much better in Emperor and it is much easier to change the colors and scale the points than in KiNG. Changing these values in KiNG requires passing a different prefs file with these modifications; however this becomes tedious with a large number of points.

9. Statistical Analyses
QIIME offers support for downstream statistical analyses. The documentation and the QIIME forum often contain discussions on the relative utility of each test and how it can be applied to microbial data. Many of the tests are implemented in R and wrapped with QIIME, so the R documentation may also be useful in interpreting the results.

a. OTU Category Significance
This script allows you to determine which OTUs are significantly associated with a metadata category or a range of metadata values. ANOVA and Pearson correlation tend to be the most useful; however this script contains a number of other tests. Make sure to use the clean OTU table, or the OTU table that resulted from filtering low abundance OTUs and single rarefaction. Since we have mostly continuous variables, the Pearson correlation for conductivity is shown below.
NOTE: This script does not work in QIIME 1.6.0. Use QIIME 1.5.0, 1.6-dev or 1.7.0.
http://qiime.org/scripts/otu_category_significance.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ otu_category_signif
icance.py -i otu_table_min10_even1000.biom -s correlation -m mapping_files/mappi
ng_complete.txt -c CONDUCTIVITY -o pearson_conductivity.txt
```

otu_category_significance.py
-i <clean_otu_table.biom>
-o <output_filepath.txt>
-m <mapping_file.txt>
-s ANOVA **or** correlation
-c <metadata_category>
ANOVA requires categorical data, Pearson requires continuous data (no NAs etc.)

The output of this script is a tab delimited file with the OTU, the p values, and the correlation coefficients. Open it in Excel for easier manipulation.

OTU	prob	oti	cat_	Bonferroni	FDR_corr	r	Consensus Lineage
913174	1.10E-08	[0.	[27(	6.29E-06	6.29E-06	0.8254	k__Bacteria; p__Bacteroidetes; c__Cytophagia; o__Cytophagales; f__Cytophagaceae; g__Leadb
4477999	9.14E-08	[0.	[27(	5.21E-05	2.61E-05	0.7951	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium;
327694	1.09E-07	[0.	[27(	6.21E-05	2.07E-05	-0.792	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudom

Figure 17. Computing Pearson correlations results in this sample output. The correlation coefficient is given by "r," while the Bonferroni and FDR corrected p-values offer two different methods of correcting p-value for multiple comparisons. The name of the OTU is listed under "Consensus Lineage."

The OTU category lists the OTU number identifier, which corresponds to the OTU name given in the Consensus Lineage column. The Bonferroni and the FDR corrected p values refer to the p value after it is corrected for multiple comparisons. Bonferroni is more conservative than FDR (false discovery rate). The r column gives the Pearson correlation of the metadata category to the OTU. In this case, the genus Leadbetterella (name cut off) is correlated strongly (r = 0.825) with conductivity, as well as *Clostridium bowmanii* (r = 0.795, name cut off). *Pseudomonas fragi* is strongly negatively correlated with conductivity (r = -0.792, name cut off). The category "prob" lists the probability that the OTU relative abundance is correlated with the category values across samples. The categories "otu_values_y" and "otu_values_x" list the data used to calculate the correlation.

b. Compare Categories
This script implements a number statistical tests to compare distance matrices, and is therefore a good way to tell how significant the clustering observed in the PCoA plot is. Below are the descriptions from selected methods.

adonis - Partitions a distance matrix among sources of variation in order to describe the strength and significance that a **categorical or continuous variable** has in determining variation of distances. This is a nonparametric method and is nearly equivalent to db-RDA (see below) except when distance matrices constructed with semi-metric or non-metric dissimilarities are provided, which may result in negative eigenvalues. adonis is very similar to PERMANOVA, though it is more robust in that it can accept either categorical or continuous variables in the metadata mapping file, while PERMANOVA can only accept categorical variables. See vegan::adonis for more details.

ANOSIM - Tests whether two or more groups of samples are significantly different based on a **categorical variable** found in the metadata mapping file. You can specify a category in the metadata mapping file to separate samples into groups and then test whether there are significant differences between those groups. For example, you might test whether 'Control' samples are significantly different from 'Fast' samples. Since ANOSIM is nonparametric, significance is determined through permutations. See vegan::anosim for more details.

PERMANOVA - This method is **very similar to adonis** except that it only accepts a **categorical variable** in the metadata mapping file. It uses an ANOVA experimental design and returns a pseudo-F value and a p-value. Since PERMANOVA is nonparametric, significance is determined through permutations.

http://qiime.org/scripts/compare_categories.html
Since we have most continuous data, the script will be shown with adonis over the metadata variable conductivity. With this script, we are determining whether or not there is a statistically significant difference between the unweighted Unifrac distances across a conductivity spectrum.

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ compare_categories.
py --method adonis -i beta_div_dms/unweighted_unifrac_otu_table_min10_even1000.t
xt -m mapping_files/mapping_complete.txt -c CONDUCTIVITY -o adonis_conductivity/
 -n 999
```

compare_categories.py
--method adonis OR anosim OR permanova

-i <input_distance_matrix.txt>
Use the distance matrix computed for beta diversity
-m <mapping_file.txt>
-c <metadata_category>
-o <output_dir>
-n <num_permutations>
Optional; the default number of permutations is 999.

The output of this script is a text file with the adonis results. The formatting is off in the tab delimited format, so refer to the compare categories tutorial to see how to break up the lines.
http://qiime.org/tutorials/category_comparison.html

Call:
adonis(formula = as.dist(qiime.data$distmat) ~ qiime.data$map[[opts$category]], permutations = opts$num_permutations)

Terms added sequentially (first to last)

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
qiime.data$map[[opts$category]]	1	0.5654	0.56542	3.2057	0.09954	0.001 ***
Residuals	29	5.1151	0.17638		0.90046	
Total	30	5.6805			1.00000	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The p-value of 0.001 (see Pr(>F)) indicates that at an alpha of 0.05, the grouping of samples by conductivity is statistically significant. The R^2 value indicates that approximately 10% of the variation in distances is explained by this grouping.

c. Compare Alpha Diversity

This script allows you to compare alpha diversity between samples using either a parameter or nonparametric two sample t-test.
http://qiime.org/scripts/compare_alpha_diversity.html

```
qiime@qiime-VirtualBox:~/Desktop/Shared_Folder/cso_for_hhmi$ compare_alpha_diver
sity.py -i collated_alpha/PD_whole_tree.txt -m mapping_files/mapping_complete.tx
t -c CONDUCTIVITY -o compare_alpha_pd_cond.txt -n 999 -t nonparametric -p bonfer
roni -d 1000
```

compare_alpha_diversity.py
-i <collated_alpha_div_filepath.txt>
This input is the output from collate_alpha.py
-m <mapping_file>
-c <metadata category>
-o <output_filepath>
-t nonparametric OR parametric
Optional; default is nonparametric
-n <num_permutations>
Optional; default is 999 permutations and is used only for nonparametric tests.
-p bonferroni OR fdr OR none

Here you can determine which multiple comparison correction method you want to use. Bonferroni is the most conservative method, while FDR (false discovery rate) is less conservative.

-d <depth of rarefaction to use>

This is the rarefaction depth for which you want to calculate significance. Typically we use the greatest depth.

The output of this script is a tab delimited file containing the pairwise alpha diversity comparisons. Open it in Excel for easier viewing.

	A	B	C	D	E	F	G	H
1	Group1	Group2	Group1 mean	Group1 std	Group2 mean	Group2 std	t stat	p-value
2	283	444	48.1198415	0.7579345	nan	nan	None	None
3	293	378	48.3624885	0.1453805	nan	nan	None	None
4	284	515	49.1782625	1.2804655	47.53336	0.003788	1.284607279	1
5	192.2	416	45.920576	0	44.120016	1.265789	0.821269422	1
6	270	192.8	48.084586	3.872245	46.393933	1.948764	0.390003416	1

Figure 18. Sample output from a nonparametric t-test comparing alpha diversity values. The column "p-value" shows which alpha diversity values is statistically different from each other. The "t-stat" column shows where the comparison fits in the constructed distribution.

We can see the results of the nonparametric sample in columns A and B, along with the respective means and standard deviations for each group. The t stat explains how the sample fits into the nonparametric distribution and the p-value indicates if the comparison is significant. In this case, none of the conductivity values were significantly different from each other.

10. Heatmaps

There is a script to generate heatmaps in QIIME; however the output is not very good. If we want to generate a heatmap of an OTU table, we use R. Below is a link to the tutorial created by Keiko Sing, a Lamendella Lab alumna and R guru, for going from OTU table to heatmap in R. Before following her instructions, make sure your OTU table is in the classic format, which has a .txt extension instead of a .biom extension. To install R, see the appendix.

convert_biom.py -i <table.biom> -o <table.from_biom.txt> -b

http://learningomics.wordpress.com/2013/02/23/from-otu-table-to-heatma/

F. QIIME Fungal ITS Workflow

1. Obtain Tutorial Files

The QIIME documentation contains example files you can use to practice the workflow. The tutorial files are from a study of the fungal community in soils Use wget to download the files and the ITS reference OTUs.

Download tutorial files and reference OTUs

```
qiime@qiime-VirtualBox:~$ wget https://s3.amazonaws.com/s3-qiime_tutorial_files/
its-soils-tutorial.tgz
```

wget https://s3.amazonaws.com/s3-qiime_tutorial_files/its-soils-tutorial.tgz

```
qiime@qiime-VirtualBox:~$ wget https://github.com/downloads/qiime/its-reference-
otus/its_12_11_otus.tar.gz
```

wget https://github.com/downloads/qiime/its-reference-otus/its_12_11_otus.tar.gz

Unzip the files using tar and gunzip.

```
qiime@qiime-VirtualBox:~/its$ tar -xzf its-soils-tutorial.tgz
qiime@qiime-VirtualBox:~/its$ tar -xzf its_12_11_otus.tar.gz
qiime@qiime-VirtualBox:~/its$ gunzip ./its_12_11_otus/rep_set/97_otus.fasta.gz
qiime@qiime-VirtualBox:~/its$ gunzip ./its_12_11_otus/taxonomy/97_otu_taxonomy.t
xt.gz
```

tar -xzf its-soils-tutorial.tgz
tar -xzf its_12_11_otus.tar.gz
gunzip ./its_12_11_otus/rep_set/97_otus.fasta.gz
gunzip ./its_12_11_otus/taxonomy/97_otu_taxonomy.txt.gz

You can see which files are in the ITS soils tutorial by going to the its-soils-tutorial directory and using ls to list the contents.

```
qiime@qiime-VirtualBox:~/its$ cd its-soils-tutorial/
qiime@qiime-VirtualBox:~/its/its-soils-tutorial$ ls
map.txt  params.txt  README.md  seqs.fna
qiime@qiime-VirtualBox:~/its/its-soils-tutorial$
```

cd /home/qiime/its-soils-tutorial
ls

The tutorial includes the sequences in fasta format (seqs.fna), a mapping file (map.txt), a parameters file (params.txt), and a readme file (README.md).

2. OTU Picking

In this tutorial, we are doing open reference OTU picking, which is a compromise between rapid closed reference OTU picking, which excludes a good chunk of sequences, and slow *de novo* OTU picking, which retains most reads.

The parameters file included a simple text file that you can input to save typing out the options into the script arguments and to specify options in workflow scripts. Many other scripts can take parameters files too, but we will not use them in this workshop. More information can be found in the QIIME documentation.
http://qiime.org/documentation/qiime_parameters_files.html

Use nano to view the parameters file.

nano params.txt

Here, the options for OTU picking and beta diversity are specified. The differences between ITS and 16S analyses are in the assign_taxonomy and the beta_diversity options. We must change the default behavior of QIIME to use the fungal ITS reference database (not Greengenes) and use the Bray-Curtis distance metric to compute beta diversity, since the phylogenetic trees needed to use Unifrac have not yet been developed for ITS.

Use ^x (control+x) to return to the command line.

Open Reference OTU Picking
We input the sequences file path, the reference ITS OTUs with (97%), define an output directory for the picked OTUs, the parameters file discussed above, and suppress aligning and constructing a phylogenetic tree from the sequences.
This step may take a half hour or more.

```
qiime@qiime-VirtualBox:~/its/its-soils-tutorial$ pick_open_reference_otus.py -i
seqs.fna -r /home/qiime/its/its_12_11_otus/rep_set/97_otus.fasta -o open_ref_its
_otus/ -p params.txt --suppress_align_and_tree
```

pick_open_reference_otus.py –i seqs.fna –r /home/qiime/its_12_11_otus/rep_set/97_otus.fasta –o open_ref_its_otus/ -p params.txt –-suppress_align_and_tree

Once the OTU table is generated, the downstream analyses in QIIME can be used to analyze the data (see the 16S workflow). Any metric that uses a phylogenetic tree (Unifrac for beta diversity and PD whole tree for alpha diversity) cannot be used with fungal ITS data at this time, since the appropriate reference trees are not yet available.

Assessments

In the Environmental Genomics folder on your flashdrive and on the Wiki you will find all of the bioinformatics and biostatiscal worksheets, quizzes, activities, discussion questions, papers, and assessments. See weeks 6 through 14. Assessments questions are also embedded in module 2 text, as well as the following questions:

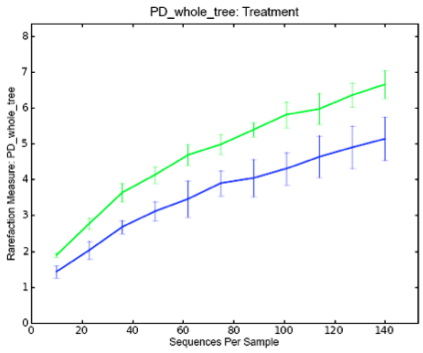
Assessment for Module 2

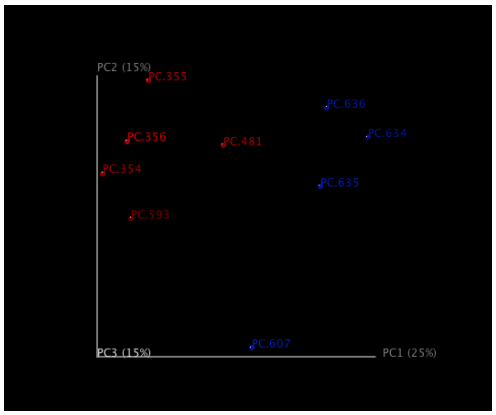
1. Briefly describe each of the steps involved in analyzing high-throughput 16S rRNA gene sequencing datasets using the qiime pipeline.

2. Describe what data one can find in an OTU table.

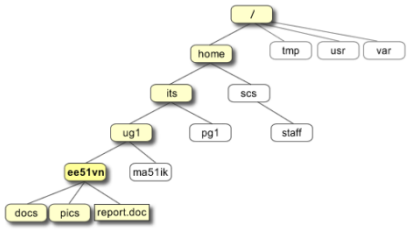
3. Compare and contrast alpha and beta diversity.

4. The plot below shows rarefaction curves for two samples green (top line) and blue (bottom line) for a measured alpha diversity metric (PD whole tree, a phylogenetic distance measure) Describe the purpose of performing rarefaction analyses and describe what the results below show about these two different samples.





5. The above plot is a Principal Coordinate Analysis based on a beta diversity metric from nine different samples. Describe the utility of these types of plots for microbial ecology research.



1. In the terminal what would you type to get to the 'pg1' directory. (Assume you start in home directory).

SampleID	BarcodeSequence	LinkerPrimerSequence	PRIMER	BARCODE
Day1.2143.A1	TCCTTAGAAGGC	GTGCCAGCMGCCGCGGTAA	H09	806 rcbc 572
Day1.2143.A2	GATGGACTTCAA	GTGCCAGCMGCCGCGGTAA	H10	806 rcbc 573
Day1.2143.B1	GGTACCTGCAAT	GTGCCAGCMGCCGCGGTAA	F04	806 rcbc 543
Day1.2143.B2	TCGCCTATAAGG	GTGCCAGCMGCCGCGGTAA	F05	806 rcbc 544
Day1.2143.C1	TCTAGCCTGGCA	GTGCCAGCMGCCGCGGTAA	D11	806 rcbc 526
Day1.2143.C2	GCCGGTACTCTA	GTGCCAGCMGCCGCGGTAA	E11	806 rcbc 538
Day2.0925.A1	CGAATGAGTCAT	GTGCCAGCMGCCGCGGTAA	E01	806 rcbc 528
Day2.0925.A2	CGATATCAGTAG	GTGCCAGCMGCCGCGGTAA	F01	806 rcbc 540

2. Identify the two problems with the above mapping file.

3. Explain in words what the following qiime python script will do. Make sure you describe each of the parameters. add_qiime_labels.py -m mapping_file.txt -i file.fasta -n 1000000 -o labelled_seqs

4. Briefly describe how we quality filtered our sequences in qiime and why its important to quality filter sequence data before performing data analysis.

5. Describe how you plan to assess microbial diversity in your study? What hypotheses do you have with respect to potential changes in diversity over the course of this study?

6. Give one advantage and one disadvantage of using a rarified OTU table to a specific sequencing depth.

7. Describe what the following qiime script does. Please include a description of the –s and –c parameters. Also describe what the expected output would look like.

otu_category_significance.py -i rarefied_otu_tables -m Map.txt -s correlation -c pH -o correlation.txt

8. Why are divergence-based measures more robust than species based beta diversity measures?

9. Describe stepwise, how you would statistically compare alpha diversity metrics between communities?

Classroom applications

Weeks 6-14 of the Environmental Genomics Research Course address the following objectives:

- Utilize a virtual machine to employ linux operating systems.
- Be able to understand directory structure, files and processes in linux
- Understand the overall workflow we will execute in qiime.
- Describe the difference between alpha and beta diversity and how we use these metrics to describe the microbial ecology of various ecosystems
- Understand how demultiplexing and quality filtering is achieved.
- Use ordination methods to compare microbial community data from different samples.
- Prepare metadata mapping files that are readable by QIIME.
- Describe the three ways OTU picking can be performed.
- Apply rarefaction analyses to normalize for sequencing depth.
- Describe methods of quantifying system change
- Describe sources of measurement error in biodiversity data and propose ways to deal with these biases.
- Be able to describe and implement both quantitative and qualitative measures of alpha and beta
- Be able to implement qiime scripts to measure alpha diversity in our samples and to statistically compare these metrics across samples.
- diversity.
- Understand the pros and cons of using phylogenetic divergence as a measure of biodiversity.

Time line of module

This module will be spread out during down times during lab activities during this 2.5 day workshop. However, in the environmental genomics research course I taught last semester we spread this out over weeks 6 through 14. Again this was my first time teaching the course so take things with a ‘grain of salt’.

Discussion Topics for class •

- The utility of Unix/Linux operating systems.

- How to develop independence to troubleshoot error messages
- Differences between alpha and beta diversity and which metric are appropriate for which biological questions
- How to choose appropriate biostatistics for your biological question
- Limitations of currently available informatics and statistical methods.

References and Suggested Reading

Diversity

1. Lozupone CA, Knight R. 2008. Species divergence and the measurement of microbial diversity. FEMS Microbiol. Rev. 32:557–578.

QIIME Algorithms

2. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26:266–267.
3. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

QIIME Examples

4. Caporaso JG, et al. 2011. Moving pictures of the human microbiome. Genome Biol. 12:R50.
5. Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. 2012.Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Microbiol. Chapter 1:Unit 1E.5.. doi: 10.1002/9780471729259.mc01e05s27.
6. QIIME allows analysis of high-throughput community sequencing data.Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. Nat Methods. 7(5):335-6.

Other useful tools

Qiime forum. Search here for helpful answers to your questions. If you have searched the forum thoroughly and have not found helpful information then post your question. The Knight group will get back to you within a few hours.
http://groups.google.com/group/qiime-forum

http://squamules.blogspot.com/2013/02/pandaseq-to-qiime.html Pandaseq to qiime. (integrating paired end sequence data and getting it into qiime.

Mothur is another 16S rRNA gene tool developed by Pat Schloss http://www.mothur.org/

AXIOME is a new tool that integrates many different tools including QIIME. Lynch MDj, Masella AP, Hall MW, Bartram AK, Neufeld JD. 2013. AXIOME: automated exploration of microbial diversity. Gigascience. doi: 10.1186/2047-217X-2-3.

APPENDIX

A. Helpful Links

QIIME links

1. Main website http://qiime.org/index.html
2. Documentation http://qiime.org/documentation/index.html
3. Scripts http://qiime.org/scripts/index.html
4. More extensive information, news, and code can be found at QIIME's github site.
   https://github.com/qiime/qiime
5. Video Tutorial for installing the QIIME VirtualBox image
   http://qiime.org/install/virtual_box.html#virtualbox-help-video

QIIME forum
https://groups.google.com/forum/?fromgroups#!forum/qiime-forum
The forum is an excellent resource. Posts range from general questions about how the scripts work to troubleshooting errors. The QIIME developers are very helpful and patient with people with limited computer science backgrounds.

Greengenes
http://greengenes.lbl.gov/cgi-bin/nph-index.cgi

Biom Format Documentation (OTU tables)
http://biom-format.org/

Linux for Beginners
http://www.linux.org/tutorial/view/beginners-level-course

Collection of File Management Commands
http://www.tuxfiles.org/linuxhelp/files.html

B. Additional Protocols/Scripts

1. Purification by SPRI Beads

This protocol allows you to remove primers from the PCR product using SPRI beads. It can be used to improve the quality of the PCR product if no gel extraction and purification is to be done.

a) Gently shake the SPRI beads bottle to resuspend any magnetic particles that may have settled.

b) Pipette 50ul of PCR product were transferred to single 96 well plate.

c) Add 90ul of SPRI beads to the samples.

d) Mix reagent and PCR reaction thoroughly by pipette mixing (very gently) and incubate mixed samples for 10 minutes at room temperature for maximum recovery. Making sure to tap gently every 2-3 minutes.

Note: this step binds PCR products 150bp and larger to the magnetic beads. The color of the mixture should appear homogenous after mixing.

e) Place the plate onto the magnetic plate for 2 minutes to separate beads from the solution.

Note: wait for the solution to clear before proceeding to the next step. At this point there will be a brown ring around the side of the tube. These are the SPRI beads containing the PCR product.

f) Remove supernatant and discard. When removing supernatant, place the pipette tip at the bottom of the well, making sure not to disturb the two bead rings.

Note: This step must be performed while the tubes are situated on the magnetic plate. Do not disturb the ring of separated magnetic beads. If beads are drawn out, leave a few microlitres of supernatant behind.

g) Add 120µl 80% ethanol to the 0.2 ml tube on the plate and incubate for 1 minute at room temperature. Pipette off the ethanol and discard. Repeat for a total of 2 washes.

Note: It is important to perform these steps with the tubes situated on the magnetic plate. Do not disturb the separated magnetic beads. Be sure to remove all the ethanol from the bottom of the tube as it is known as PCR inhibitor.

h) Let excess alcohol evaporate (≤ 5 minutes).

Note: Take care not to over dry the bead ring (bead ring appears cracked) as this will significantly decrease elution efficiency.

i) Remove the tubes from the plate. Add 40µl of elution buffer (1x TE) and mix by pipetting. This will separate the beads from the PCR product.

Note: The liquid level needs to be high enough to contact the magnetic beads. A greater volume of elution buffer can be used, but using a lower volume might require extra mixing and may not fully elute the entire product. Elution is quite rapid and it is not necessary for the beads to go back into solution for it to occur.

I would add 40 µl if the samples were previously concentrated in a final volume of 50 µl, or I would add 50 µl if the samples were for the PCR replicates combined in a final volume of 75 µl.

j) Place the tubes back onto the magnetic plate for 1 minute to separate the beads from the solution. Transfer the eluent to a new tube. This contains the purified PCR product.

Pool PCR amplicons in equimolar concentrations

a) Quantify the amplicon pools by electrophoresis or bioanalyzer (preferably by bioanalyzer).

b) Pool 50 ng of each sample together in a 1.5 ml tube. At this stage, the volume is unimportant.

Note: Lower quantities of each sample can be pool together. However, pooling a minimum amount of 50 ng of each sample, will assure the final amount of 2 μg total that sequencing requires.

Concentration (Option A)

The pooled samples will need to be concentrated into 25μl in order for the sequencing libraries to be created.

a) Combine the triplicates PCR products in a thin-walled 0.2 ml tube.

b) Add 1 μl of linear acrylamide to pooled products.

Note: this acts as a carrier for small amounts of DNA.

c) Add 1/10 volume of sodium acetate.

d) Add 0.8 - 1 volume of isopropanol.

e) Mix well by pipetting.

f) Incubate at -80 °C for approximately 15 minutes.

g) Centrifuge at maximum speed at 4 °C for 20 minutes.

h) Remove the supernatant.

i) Wash the pellet with 1 volume of 70% ethanol.

Note: 70% ethanol is hygroscopic. Fresh 70% ethanol should be prepared for optimal results.

j) Centrifuge again at maximum speed at 4 °C for 20 minutes.

k) Remove the supernatant.

l) Air-dry the pellet for 3-4 minutes.

Note: Tubes can be placed in water bath at 37 ºC for 5 minutes.

m) Resuspend the pellet in 25μl. TE buffer **(25ul of 2ug for half plate)**

Concentration (Option B)

The pooled samples will need to be concentrated into 25μl in order for the sequencing libraries to be created.

Using Millipore Centrifugal filter Units (Amicon Ultra 0.5ml 30K membrane). Pool samples from half plate together before concentration.

1. Add ~40ul of TE/water to the filter, spin the filter at 14,000g for 5minutes.
    a. This is a pre-wet step to get better yields
2. Add the sample to the filter and centrifuge at 14,000g at room temperature for 8minutes
    a. Time varies depending on the starting volume
3. Final volume left in the filter unit should be about 25ul.
4. Flip the filter unit and re-centrifuge at 1000g for 2minutes.

2. Splitting Libraries – The traditional method

Often, MiSeq sequencing platforms demultiplex the data with the files in fastq format, which does not fit well into the QIIME workflow. Typically, splitting the libraries is done with barcoded fastq files, with separate files containing the barcodes. These are then input together, and the script demultiplexes the reads, assigns them a unique identifier, quality filters, and converts the sequences to fasta format for OTU picking.

We can still use the QIIME script with a few modifications.
http://qiime.org/scripts/split_libraries_fastq.html

split_libraries_fastq.py
-i <sequence_read_fastq_file.fastq>
    Separate these by commas (no spaces) if there is more than one.
-o <output_dir>
-m <mapping_file>
    Separate these by commas (no spaces) if there is more than one.
--sample_id <file_name_when_using_demultiplexed_data>
    Use only when the sequences are demulitplexed (in our case)
--rev_comp
    This is required if you are using reverse reads.
-q <minimum_phred_score>
    This is the minimum Phred Score a base may have. A Phred score of 20 corresponds to a 99% chance that the base was called correctly.
--barcode_type 'not-barcoded'

C. Other Software

1. Other ways to analyze metagenomes

mothur http://www.mothur.org/
Picante R package http://picante.r-forge.r-project.org/

2. Installing R

R is already installed in the QIIME VDI. This tutorial covers installing R on your Windows machine. This tutorial is by Keiko Sing and can be found along with an introduction to R at her blog.

http://learningomics.wordpress.com/2013/01/28/i-thought-r-was-a-letter-an-introduction/

a. Go to the R website and click **"Download R"** under "Getting Started"

b. Choose a place to download R. Choosing a location close to you helps speeds things up.

c. Choose which R package to download based on your operating system in the first box. If you are Unix or Mac user, I apologize but this is where we now go our separate ways.

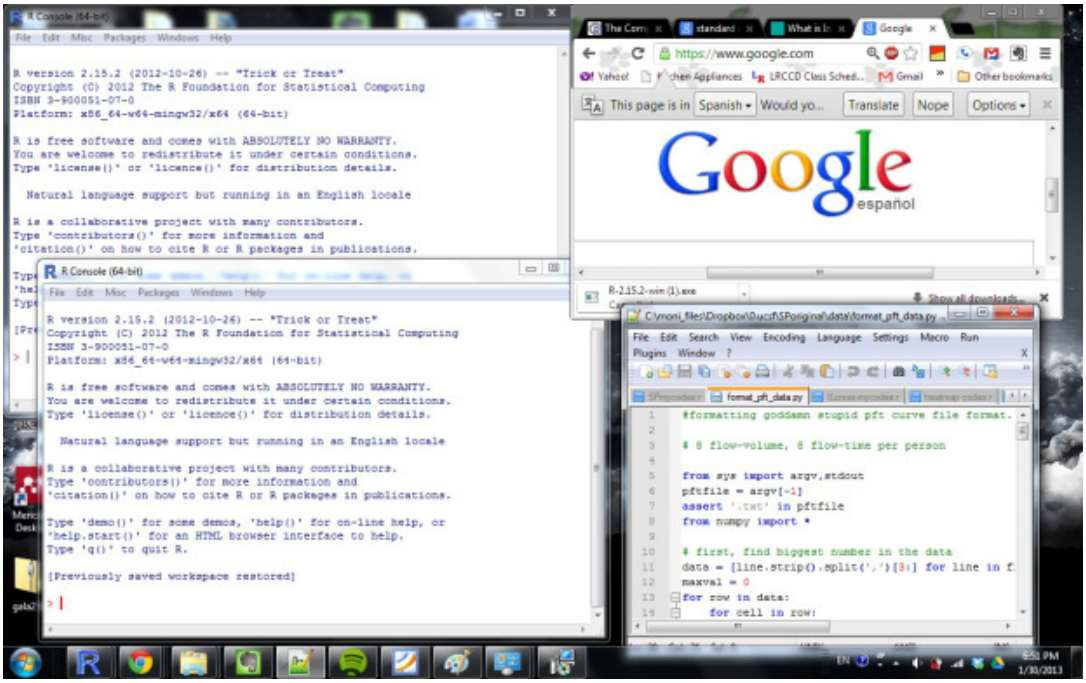
d. Click on **"install R for the first time"** then download the file with the biggest font on the top.

e. Click **"run"**. Then choose your language.

f. Click **"next"** to start the installation, agree to all their legal writings, and selection an installation window.

g. Select "Core Files" and then either 32-bit or 64-bit files depending on your computer system. (To check, hit **Start**, right click **Computer** and select **Properties**. Look at **System Type**).

h. Now you have a choice for Startup Options. I prefer to view the program in multiple separate windows so that I can arrange them on my screen while also have an internet browser or a notepad type program open as well.


R Console (64-bit)
R version 2.15.2 (2012-10-26) -- "Trick or Treat"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Previously saved workspace restored]
https://www.google.com
Google
español
#formatting goddamn stupid pft curve file format.
# 8 flow-volume, 8 flow-time per person
from sys import argv,stdout
pftfile = argv[-1]
assert '.txt' in pftfile
from numpy import *
# first, find biggest number in the data
maxval = 0
for row in data:
for cell in row:

If you like what you see in the photo above, click **"Yes (customized setup)"**. If you prefer to have one window with all the components of the program viewed inside that window click **"No (accept defaults)"** and skip to **Step 11**.

i. If you said yes to **Step 8**, click **"SDI (separate windows)"**. Next, you can specify plain text or HTML help. I would suggest HTML help because it is easier to view than plain text, which appears in the window.

j. If you are at an institution that utilizes Internet2.dll, select **"Internet 2."** If not or if you are unsure, select **"Standard".**

k. Go ahead and create a program shortcut by clicking **"Next**".

l. Choose if you want to have another icon clutter your desktop and/or Quick Launch toolbar. I suggest leaving the two options under **"Registry Entries"** selected.

3. Proprietary software for data analysis
These tools were originally designed for macroecologists, but can be used with microbial ecology data as well. These software packages are not free, unlike QIIME and R.

a. PC-ORD is a software package for multivariate community analysis.
http://home.centurytel.net/~mjm/pcordwin.htm

b. Primer E with the PERMANOVA+ add-on is another software package for multivariate community statistics. It is the premier software package used in ecology, and includes many analyses QIIME does not yet implement, including distance-based redundancy analysis (db-RDA). The software comes with an excellent manual, which describes in detail each statistical analysis and how to use them to analyze your communities.
http://www.primer-e.com/

4. Computing

a. Troubleshooting QIIME error messages
There are three general types of error messages you will encounter when working with QIIME. They include shell error messages, QIIME-specific error messages, and Python error messages. Shell error messages typically diminish as you become more comfortable working with the command line. Python error messages are typically the hardest to troubleshoot, since they typically signal some sort of formatting issue, which may or may not be apparent to the user. Sometimes you will get no error message, but a script may hang without completing.

Shell Error Messages
No such file or directory
You are probably not looking for a file or directory in the appropriate place, or you are attempting to pass a file into a script that is not found in the path you have provided. Check the paths to make sure the files are indeed found there. Use tab as much as possible to avoid these error messages, since the shell will not let you input something it cannot interpret. Check spelling and capitalization too. The shell is case sensitive, so Shared_Folder and Shared_folder are two different directories. Avoid spaces and try to keep names informative but short to limit this type of error.

Permission denied
VirtualBox will not let you write directly to the Shared_Folder. You can write to subdirectory of the Shared_Folder, so on your host machine (the laptop or whatever machine VB is installed on), create a subfolder where you are able to write files or directories. If this is not a

Shared_Folder issue, use chmod to change the permissions so you may write or execute certain directories or processes.

QIIME-specific Error Messages
Usually QIIME-specific error messages will return the script you attempted to run, plus a list of the options the script can take. Each option will specify which kinds of files it takes (filepath or directory) and what sort of data should be included in the input files. If you are still stuck after carefully checking the script input against the documentation, check the QIIME help forum. It is likely that several other people have run into the same issue.

Python Error Messages
Python error messages are the most cryptic error messages. Usually, formatting issues or bugs in the scripts cause these sorts of errors. They are general, so other people using Python (a programming language used in scripting) may have encountered them before. Typically, they involve Python expecting a string (of characters) instead of a float (number with a decimal) or vice versa. Sometimes there are temporary files left in directories even after scripts have completed that Python will complain about. In previous versions of QIIME, OTU tables in BIOM format were typically suspect to Python errors because of Consensus Lineage formatting issues in earlier versions of BIOM. If you run into Python errors, check the QIIME forum or consult other resources like Stack Overflow to troubleshoot the error messages. Occasionally, there is a bug in the script, reports of which you can find on the QIIME Github site.

Hanging Scripts
When you are dealing scripts that require a lot of memory, such as open reference or *de novo* OTU picking or multiple rarefactions, sometimes the script hangs without completing. From the command line, it appears that the script is still running, but the memory usage and the CPU usage (see the System Monitor) are low. There might be a temporary file that did not get deleted and the script will therefore not proceed any further. It could also be that the script is performing very inefficiently (swap is high), and will not complete in a reasonable time frame. Some worfklow scripts, like alpha_rarefaction.py and any of the OTU picking workflows are typically suspect to "hanging errors." In these cases attempt to split up the workflow so that you have better control over the steps. You may need to upgrade RAM, use more resources on a cluster, or check out the elastic compute cloud from Amazon Web Services to get enough RAM to complete the script.

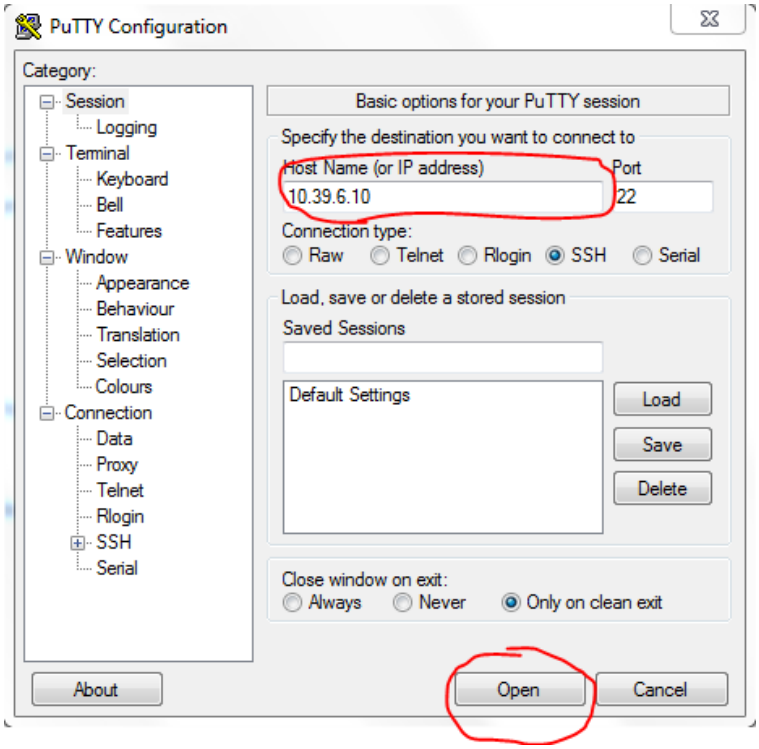
2. Connecting to Juniata's HHMI Cluster
This guide by Alex Sickler is a detailed protocol for connecting remotely to the cluster for both Unix/Linux and Windows users.

Mac or Linux:

open a terminal window
enter ssh username@10.39.6.10

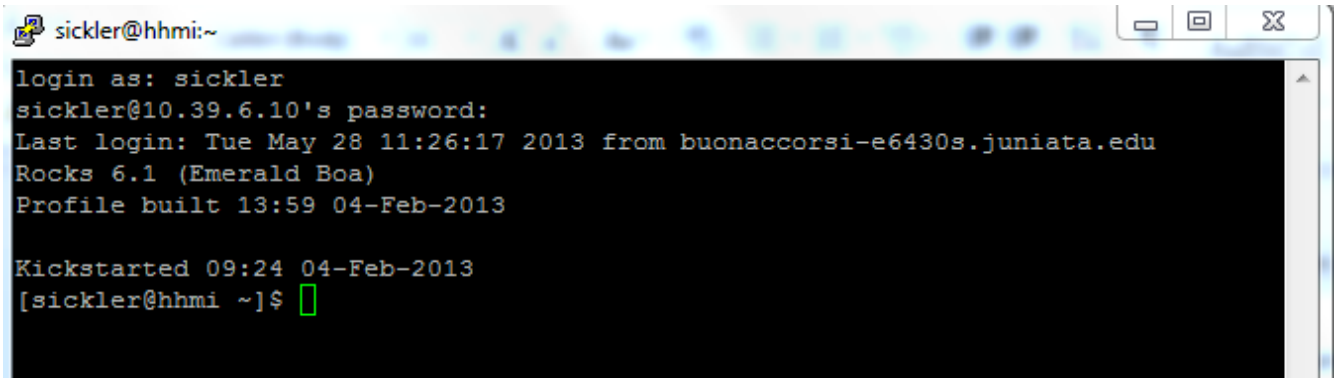
Windows:

Download putty from
http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
Launch putty
In the host name box enter 10.39.6.10
Click open



A terminal window will open asking for your username and password.
If you are asked about the rsa2 key, click yes.

All Operating Systems:
Once you are logged in, you should see a message similar to the one below and a command prompt:



Configuring your bashrc file:
To use the programs that are installed on our cluster, you first need to set your PATH variable which tells the computer where to look for programs.

To configure your bashrc file with the path to the programs run the following commands:
cp ~sickler/.bashrc ./.bashrc
source .bashrc

To verify everything worked enter:
which nseg

the computer should return:

/share/apps/nseg/nseg

3. IPython Notebook

Several QIIME tutorials make use of IPython Notebook, which is a tool that combines code with plain text and figures. IPython Notebook is viewed in a web browser, and is accessible as a static page to everyone you share it with, even if they do not have IPython Notebook installed themselves. IPython Notebook can be used to build workflows, demonstrate scripts in realtime, or keep lab notebooks for labs that are more computationally-minded.

IPython Notebook is easy to install and fun to play around with. There are a number of sample galleries, as well as instructions on how to get the most out of this tool.

1. IPython Notebook website
2. IPython viewer
3. Cool Notebooks
4. Learning Python with IPython Notebook
5. Quick Installation Guide

QIIME & IPython Notebook

Illumina Overview IPython Notebook Tutorial

Fungal ITS IPython Notebook Tutorial

Integrating QIIME and IPython Notebook

4. Bash Scripting

Bash Scripting can be used to automate repetitive scripts or tasks that require lots of manual file management. While you may be tempted to physically bash the computer, correctly implemented bash scripts can save a lot of time and reduce human error.

The typical bash scripts we use with QIIME are loops to automate such tasks as file conversion or file truncation. Several bash scripts are used in this tutorial, and they are based on reading a file line by line and performing some task while there are still lines to be read in the file.
Make a list of the files you want to do something to by using ls and writing the list to a file.

ls > list.txt

The general architecture of a while loop is below. While the first line is read, do the script, and close the list once all the lines have been read. Time is there to timestamp the scripts (but is not required), which may be helpful for estimating future runtimes and troubleshooting memory intensive scripts.

```
while read line;
do time <copy and paste your script here, using $line as the input>;
done < list.txt
```

Beginner's Guide to Bash Scripting
http://www.tldp.org/LDP/Bash-Beginners-Guide/html/

Short History and Overview of Bash
https://en.wikipedia.org/wiki/Bash_%28Unix_shell%29

E. References

1. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1,** 4516–4522 (2011).
2. Gardes, M. & Bruns, T. D. ITS primers with enhanced specificity for basidiomycetes--application to the identification of mycorrhizae and rusts. *Mol. Ecol.* **2,** 113–118 (1993).
3. McGuire, K. L. *et al.* Digging the New York City Skyline: Soil Fungal Communities in Green Roofs and City Parks. *Plos One* **8,** e58020 (2013).
4. Qubit 2.0 User Manual http://www.invitrogen.com/etc/medialib/en/filelibrary/cell_tissue_analysis/Qubit-all-file-types.Par.0519.File.dat/Qubit-2-Fluorometer-User-Manual.pdf
5. Qubit dsDNA HS Manual. http://probes.invitrogen.com/media/pis/mp32851.pdf
6. Bioanalyzer DNA Assay 1000 Manual http://www.chem.agilent.com/library/usermanuals/Public/G2938-90014_KitGuideDNA1000Assay_ebook.pdf
7. Bioanalyzer DNA Assay 1000 Quick Protocol

http://www.chem.agilent.com/library/usermanuals/Public/G2938-90015_QuickDNA1000.pdf